

OrderPatent

{19}

JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: 06332632 A

(43) Date of publication of application: **02.12.1994**

(51) Int. Cl. G06F 3/06

G06F 3/06, G06F 11/10

(21) Application number: 05125766

(22) Date of filing: 27.05.1993

(71) Applicant: **HITACHI LTD**

(72) Inventor: TSUNODA HITOSHI
TAKAMOTO YOSHIFUMI

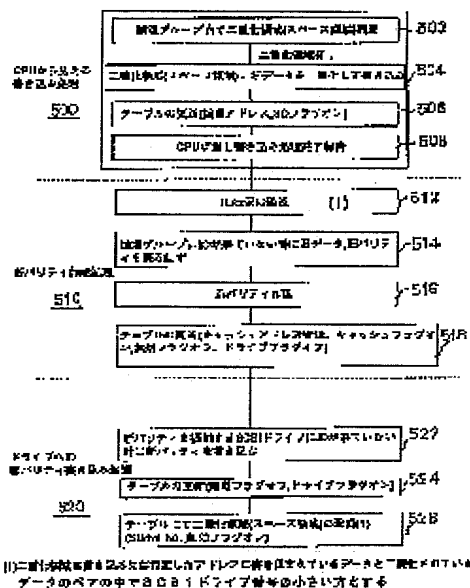
(54) DISK ARRAY DEVICE AND ITS CONTROL METHOD

COPYRIGHT: (C)1994.JPO

(57) Abstract:

PURPOSE: To decrease the overhead at the time of writing data, in the disk array system of an RAID (level 5) for improving the processing performance by distributing the data.

CONSTITUTION: In a group for generating the parity, a duplex area is provided. At the time of write, data to be written in is written in the duplex area first (step 502), and at that time point, a write processing to a CPU is finished (508). The parity generation is executed at the subsequent suitable timing (516), and the data is written in a SCSI drive (522). That is, the generation of the parity is efficiently scheduled separately from write of the data. At the time of write of the data, the data to be written is written in the duplex area, and by processing independently timewise write of the data to the SCSI drive and the generation of the parity, like this, the parity generation overhead at the time of write becomes invisible from a CPU.



(51) Int.Cl. ⁵	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 3/06	3 0 5 C			
	3 0 1 Z			
11/10	3 2 0 Z			

審査請求 未請求 請求項の数28 O L (全 23 頁)

(21) 出願番号 特願平5-125766

(22) 出願日 平成5年(1993)5月27日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 角田 仁

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(72) 発明者 高本 良史

東京都国分寺市東恋ヶ窪1丁目280番地

株式会社日立製作所中央研究所内

(74) 代理人 弁理士 薄田 利幸

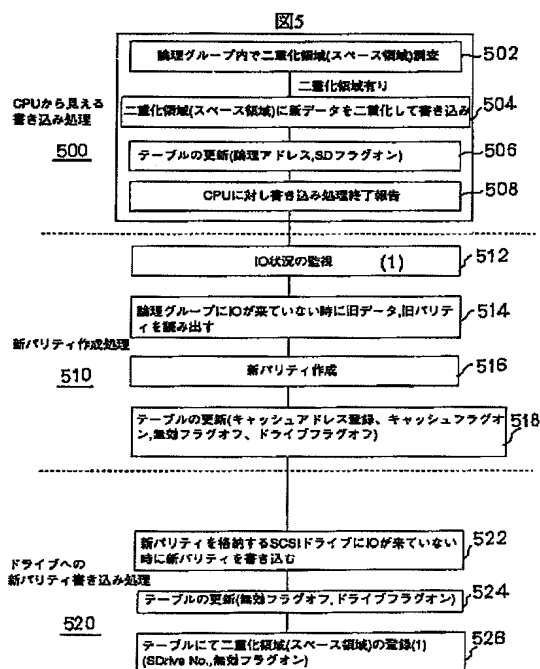
(54) 【発明の名称】 ディスクアレイ装置及びその制御方法

(57) 【要約】

【目的】データを分散させて処理性能を向上させる R A I D (レベル5) のディスクアレイシステムにおいて、データの書き込み時のオーバーヘッドを減少させる。

【構成】パリティを作成されるグループにおいて、二重化領域を設ける。書き込み時において、書き込むデータを二重化領域にとりあえず書き込み (ステップ502)、その時点でCPUに対し書き込み処理を終了とする (508)。パリティ作成は後の適当なタイミング (516) で行い、S C S I ドライブに書き込む (522)。つまり、パリティの作成をデータの書き込みとは別に、効率良くスケジューリングする。

【効果】データの書き込み時において、書き込むデータを二重化領域に書き込み、このように、データの S C S I ドライブへの書き込みと、パリティの作成を、時間的に独立に処理することにより、書き込み時のパリティ作成オーバーヘッドはCPUには見えなくなる。



(1) 二重化領域は書き込み先に指定したアドレスに書き込まれているデータと二重化されているデータのペアの中で S C S I ドライブ番号の小さい方とする

【特許請求の範囲】

【請求項1】データを格納する複数のドライブ群と、該ドライブ群を管理する制御装置とを備え、書き込みデータを前記ドライブ群に格納し、該各ドライブに格納されているデータにより誤り訂正符号を生成し、この生成した誤り訂正符号を、該誤り訂正符号を生成するのに関与したデータが格納されているドライブとは別のドライブに格納する記憶装置の制御方法において、データ書き込みに伴う誤り訂正符号の更新を、前記データ書き込み処理よりも遅延させて処理することを特徴とする記憶装置の制御方法。

【請求項2】書き込むべき複数のデータから少なくとも1つの誤り訂正符号を含む1つの論理グループを生成し、該論理グループを構成する複数のデータ及び誤り訂正符号を複数のドライブ群内に格納する記憶装置におけるデータの格納方法であって、前記論理グループを構成する複数のデータの1つを新たなデータに書き換える更新要求に応答して、該更新データと新たな誤り訂正符号とを含む1つの論理グループを新たに生成し、前記新たな論理グループの更新データと誤り訂正符号とを、前記書き換え前のデータ及び誤り訂正符号が格納されていたドライブとは異なるドライブにそれぞれ分散して格納する記憶装置におけるデータの格納方法において、データ書き込みに伴う誤り訂正符号の更新を、前記データ書き込み処理よりも遅延させて処理することを特徴とする記憶装置の制御方法。

【請求項3】データを格納する複数のドライブ群と、該ドライブ群を管理する制御装置とを備え、書き込みデータを前記ドライブ群に格納し、該各ドライブに格納されているデータにより誤り訂正符号を生成し、この生成した誤り訂正符号を、該誤り訂正符号を生成するのに関与したデータが格納されているドライブとは別のドライブに格納する記憶装置の制御方法において、データ書き込みに伴う誤り訂正符号の更新を、データの書き込みを要求した上位置装置の負荷が所定値以下のときに処理することを特徴とする記憶装置の制御方法。

【請求項4】書き込むべき複数のデータから少なくとも1つの誤り訂正符号を含む1つの論理グループを生成し、該論理グループを構成する複数のデータ及び誤り訂正符号を複数のドライブ群内に格納する記憶装置におけるデータの格納方法であって、前記論理グループを構成する複数のデータの1つを新たなデータに書き換える更新要求に応答して、該更新データと新たな誤り訂正符号とを含む1つの論理グループを新たに生成し、前記新たな論理グループの更新データと誤り訂正符号とを、前記書き換え前のデータ及び誤り訂正符号が格納されていたドライブとは異なるドライブにそれぞれ分散して格納する記憶装置におけるデータの格納方法において、

データ書き込みに伴う誤り訂正符号の更新を、データの書き込みを要求した上位置装置の負荷が所定値以下のときに処理することを特徴とする記憶装置の制御方法。

【請求項5】データを格納する複数のドライブ群と、該ドライブ群を管理する制御装置とを備え、書き込みデータを前記ドライブ群に格納し、該各ドライブに格納されているデータにより誤り訂正符号を生成し、この生成した誤り訂正符号を、該誤り訂正符号を生成するのに関与したデータが格納されているドライブとは別のドライブに格納するにおいて、

データ書き込みに伴う誤り訂正符号の更新を、前記制御装置の負荷が所定値以下のときに処理することを特徴とする記憶装置の制御方法。

【請求項6】書き込むべき複数のデータから少なくとも1つの誤り訂正符号を含む1つの論理グループを生成し、該論理グループを構成する複数のデータ及び誤り訂正符号を複数のドライブ群内に格納する記憶装置におけるデータの格納方法であって、

前記論理グループを構成する複数のデータの1つを新たなデータに書き換える更新要求に応答して、該更新データと新たな誤り訂正符号とを含む1つの論理グループを新たに生成し、前記新たな論理グループの更新データと誤り訂正符号とを、前記書き換え前のデータ及び誤り訂正符号が格納されていたドライブとは異なるドライブにそれぞれ分散して格納する記憶装置におけるデータの格納方法において、

データ書き込みに伴う誤り訂正符号の更新を、前記制御装置の負荷が所定値以下のときに処理することを特徴とする記憶装置の制御方法。

【請求項7】データを格納する複数のドライブ群と、該ドライブ群を管理する制御装置とを備え、書き込みデータを前記ドライブ群に格納し、該各ドライブに格納されているデータにより誤り訂正符号を生成し、この生成した誤り訂正符号を、該誤り訂正符号を生成するのに関与したデータが格納されているドライブとは別のドライブに格納するディスクアレイシステムにおいて、

データ書き込みに伴う誤り訂正符号の更新を、前記データの書き込みとは非同期でかつ所定のタイミング毎に処理することを特徴とするディスクアレイ装置の制御方法。

【請求項8】あるドライブに障害が発生し、障害が発生したドライブ内のデータまたはパリティを回復処理により復元し、この復元したデータまたはパリティを予備の書き込み領域に書き込んだ後、障害が発生したドライブを、正常なドライブに交換し、交換後は、この交換した正常なドライブは全てスペース領域により構成されているとして論理グループを再構成して処理を再開することを特徴とする請求項1記載の記憶装置の制御方法。

【請求項9】あるドライブに障害が発生したことを感知したら、障害が発生したドライブを正常なドライブに交

換し、障害が発生したドライブ内のデータまたはパリティを回復処理により復元し、この復元したデータまたはパリティと、障害が発生したドライブ内にあったスペース領域を、交換した正常なドライブに書き込んで再構成して処理を再開することを特徴とする請求項1記載の記憶装置の制御方法。

【請求項10】最新に書き込まれたデータについては二重化して高信頼とし、書き込み要求があまり発行されないデータについてはパリティにより信頼性を確保するように、信頼性について二段階のレベルを設定することを特徴とする請求項1記載の記憶装置の制御方法。

【請求項11】書き込むべき複数のデータから少なくとも1つの誤り訂正符号を含む1つの論理グループを生成し、該論理グループを構成する複数のデータ及び誤り訂正符号を複数のドライブ群内に格納する記憶装置であって、

前記論理グループを構成する複数のデータの1つを新たなデータに書き換える更新要求にตอบสนองして、該更新データと新たな誤り訂正符号とを含む1つの論理グループを新たに生成し、前記新たな論理グループの更新データと誤り訂正符号とを、前記書き換え前のデータ及び誤り訂正符号が格納されていたドライブとは異なるドライブにそれぞれ分散して格納する記憶装置において、データ書き込みに伴う誤り訂正符号の更新を、前記データ書き込み処理よりも遅延させて処理することを特徴とする記憶装置。

【請求項12】前記記憶装置を構成するドライブの集合の中に、書き込むデータを一旦二重化して書き込める領域をもつことを特徴とする請求項11記載の記憶装置。

【請求項13】スペース領域を、パリティを生成するデータと、それらのデータにより生成したパリティが格納されているドライブに分散してもつことを特徴とする請求項11記載の記憶装置。

【請求項14】スペース領域を、パリティおよびパリティの作成に関与したデータの格納されているドライブ以外の、異なる2台のドライブに確保することを特徴とする請求項11記載の記憶装置。

【請求項15】パリティを生成するデータと生成したパリティと、スペース領域が、それぞれ別のドライブに格納されており、これらの対応するデータ、パリティ、スペース領域が、各ドライブにおいて同一の物理的なアドレスに格納されるよう制御する手段を備えていることを特徴とする請求項13記載の記憶装置。

【請求項16】パリティを生成するデータと生成したパリティと、スペース領域が、それぞれ別のドライブに格納されており、これらの対応するデータ、パリティ、スペース領域は、各ドライブにおいて任意の物理的なアドレスに格納されるよう制御する手段を有することを特徴とする請求項13記載の記憶装置。

【請求項17】上位装置からの書き込み要求に対し、ディ

スク装置を制御する制御装置が、書き込みデータを、スペース領域の確保されている2台のドライブの物理的なアドレスに二重化して書き込む制御を行うプロセッサを持つことを特徴とする請求項13記載の記憶装置。

【請求項18】論理グループ内において、上位装置からの新データ書き込み要求に対し、ディスク装置を制御する制御装置は、スペース領域の確保されている2台のドライブの物理的なアドレスに、二重化して書き込む様に変換して制御するプロセッサを持つことを特徴とする請求項13記載の記憶装置。

【請求項19】上位装置から書き込み要求が発行された場合、書き込む新データをスペース領域の確保されている2台のドライブの物理的なアドレスに二重化して格納し、この段階で上位装置に対し、書き込み処理の終了を報告するプロセッサを持つことを特徴とする請求項11記載の記憶装置。

【請求項20】上位装置から書き込み要求が発行された場合、書き込む新データをスペース領域の確保されている2台のドライブの物理的なアドレスに二重化して格納し、この段階で上位装置に対し、書き込み処理の終了を報告し、データの当該ドライブへの書き込み処理とは独立に、後で書き込みによるパリティの作成および当該ドライブへの書き込みを行うように制御するプロセッサを持つことを特徴とする請求項18記載の記憶装置。

【請求項21】上位装置から書き込み要求が発行された場合、書き込む新データをスペース領域の確保されている2台のドライブの物理的なアドレスに二重化して格納し、この段階で上位装置に対し、書き込み処理の終了を報告した後、上位装置からの読み出しまたは書き込み要求数をカウントし、このカウント値をユーザまたはシステム管理者が予め設定した数とで比較する制御を行うプロセッサを持つことを特徴とする請求項19記載の記憶装置。

【請求項22】上位装置から書き込み要求が発行された場合、書き込む新データをスペース領域の確保されている2台のドライブの物理的なアドレスに二重化して格納し、この段階で上位装置に対し、書き込み処理の終了を報告した後、上位装置からの読み出しまたは書き込み要求数をカウントし、このカウント値をユーザまたはシステム管理者が予め設定した数とで比較した結果、カウント値が設定数より小さく、しかも、当該ドライブに対し上位装置から読み出しまたは書き込み要求が発行されていない場合、書き込み処理におけるパリティの作成を開始するように判断する制御を行うプロセッサを持つことを特徴とする請求項20記載の記憶装置。

【請求項23】論理グループにおいて、パリティを作成するデータとパリティの集合をパリティグループとし、書き込み前と書き込み後では、パリティグループを構成するデータおよびパリティの格納されているドライブが異なることを特徴とする請求項11記載の記憶装置。

【請求項24】論理グループ内のあるドライブに障害が発生した場合、正常な残りのドライブ内のデータとパリティから、障害が発生したドライブ内のデータまたはパリティを回復処理により復元するが、この復元したデータまたはパリティをスペース領域に書き込む制御を行うプロセッサを持つことを特徴とする請求項1記載のディスクアレイシステム。

【請求項25】書き込みにより作成した新パリティを、当該ドライブへ書き込む前に、あるドライブに障害が発生した場合、旧パリティの作成に関与したデータについて、この旧パリティの作成に関与した、正常な残りのドライブ内のデータとパリティから回復処理により復元し、二重化されている新データが格納されているドライブに障害が発生した場合は、二重化データの一方から、障害が発生したドライブ内のデータまたはパリティを回復処理により復元する制御を行うプロセッサを持つことを特徴とする請求項1記載のディスクアレイシステム。

【請求項26】書き込みにより作成した新パリティを、当該ドライブへ書き込む前に、あるドライブに障害が発生した場合、正常な残りのドライブ内のデータとパリティと二重化されている新データから、障害が発生したドライブ内のデータまたはパリティを回復処理により復元するが、この復元したデータまたはパリティをスペース領域に書き込む制御を行うプロセッサを持つことを特徴とする請求項1記載のディスクアレイシステム。

【請求項27】データとパリティとスペース領域を、テーブルにより管理することを特徴とする請求項1記載のディスクアレイシステム。

【請求項28】論理グループ単位に、その内部にアドレス変換用テーブルと、パリティ生成回路と、キャッシュメモリと、それらを制御するマイクロプロセッサを持つことを特徴とする請求項1記載のディスクアレイシステム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明はコンピュータシステムに係り、特に高性能な入出力動作を可能とするディスクファイルシステムに関する。

【0002】

【従来の技術】現在のコンピュータシステムにおいては、CPU等の上位側が必要とするデータは2次記憶装置に格納され、CPUが必要とする時に応じ2次記憶装置に対してデータの書き込み、読みだしを行っている。この2次記憶装置としては一般に不揮発な記憶媒体が使用され、代表的なものとして磁気ディスク装置（以下ドライブとする）、光ディスクなどがあげられる。

【0003】近年高度情報化に伴い、コンピュータシステムにおいて、2次記憶装置の高性能化が要求されてきた。その一つの解として、多数の比較的容量の小さなド

ライブにより構成されるディスクアレイが考えられている。

【0004】「D.Patterson,G.Gibson,and R.H.Kartz;A Case for Redundant Arrays of Inexpensive Disks(RAID),in ACM SIGMOD Conference,Chicago,IL,(June1988)」において、データを分割して並列に処理を行うディスクアレイ（レベル3）とデータを分散して、独立に扱うディスクアレイ（レベル5）について、その性能および信頼性の検討結果が報告されている。現在この論文に書かれている方式が最も一般的なディスクアレイと考えられている。

【0005】以下にデータを分散して、独立に扱うディスクアレイ（レベル5）について説明する。レベル5のディスクアレイでは個々のデータを分割せずに独立に扱い、多数の比較的容量の小さなドライブに分散して格納するものである。現在、一般に使用されている汎用大型コンピュータシステムの2次記憶装置では、1ドライブ当りの容量が大きいため、他の読み出し／書き込み要求に当該ドライブが使用されて、そのドライブを使用できずに待たされることが多く発生した。このタイプのディスクアレイでは汎用大型コンピュータシステムの2次記憶装置で使用されている大容量のドライブを、多数の比較的容量の小さなドライブで構成し、データを分散して格納してあるため、読み出し／書き込み要求が増加してもディスクアレイの複数のドライブで分散して処理するため、読み出し／書き込み要求がまたされることが減少する。しかし、ディスクアレイは、このように多数のドライブにより構成されるため、部品点数が増加し障害が発生する確率が高くなる。そこで、信頼性の向上を図る目的で、パリティを用意する必要がある。

【0006】このパリティによりデータを格納したドライブに障害が発生した場合、その障害ドライブ内のデータを復元することが可能となる。ディスクアレイではデータからパリティを作成しデータと同様にドライブに格納しておく。この時、パリティは、パリティの作成に関与したデータとは別のドライブに格納される。

【0007】これらのディスクアレイでは、現在一般に使用されている汎用大型コンピュータシステムと同様、2次記憶装置内では、個々のデータの格納場所（アドレス）は予め指定したアドレスに固定され、CPUから当該データへ読みだしまたは書き込みする場合、この固定されたアドレスへアクセスすることになっている。この分散して格納するディスクアレイ（レベル5）ではストレージテクノロジーポレーション（以下STKとする）から製品発表がされている。米国特許WO 91/20076では、レベル5の基本アーキテクチャにおいて、動的に変更可能なアドレスのテーブルを用意することにより、データ圧縮を行いデータの書き込み処理において、トラック単位で書き込み先のアドレスを動的に変換する方法について開示されている。

【0008】また、特開平4-230512にはレベル5において、書き込み時に書き込むデータと、この書き込みにより更新したパリティを、それぞれ別の場所に書き込む方法について開示されている。さらに、IBM社のディスクアレイ(9337)では、レベル5においてWAD(ライト アシスト デバイス)を設けることが発表されている。

【0009】

【発明が解決しようとする課題】現在の汎用大型計算機システム等ではドライブにより構成される2次記憶装置内では、CPUから転送されてくるデータは個々のデータの格納場所(アドレス)が予め指定したアドレスに固定され、CPUから当該データへ読みだしまたは書込む場合は、この固定されたアドレスへアクセスすることになる。これは、ディスクアレイにおいても同じである。データを分割して並列に処理を行うディスクアレイ(レベル3)ではこのようにアドレスを固定しても影響は無いが、データを分散して、独立に扱うディスクアレイ(レベル5)ではアドレスを固定した場合、書き込み時に大きな処理オーバーヘッドが必要になる。以下それについて説明する。

【0010】図11は前記公知例でD.Pattersonらが提案したRAIDに述べられている、データを分散して、独立に扱うディスクアレイ(レベル5)内部のデータアドレスを示している。この各アドレスにあるデータは1回の読み出し/書き込み処理される単位で、個々のデータは独立している。また、RAIDに述べられているアーキテクチャではデータに対するアドレスは固定されている。前述したようにこのようなシステムでは、信頼性を向上するためパリティを設定することが不可欠である。本システムでは各ドライブ内の同一アドレスのデータによりパリティが作成される。すなわち、ドライブ#1から4までのアドレス(1, 1)のデータによりパリティが作成され、パリティを格納するドライブの(1, 1)に格納される。本システムでは読み出し/書き込み処理は現在の汎用大型計算機システムと同様に各ドライブに対し当該データをアクセスする。

【0011】このようなディスクアレイにおいて、例えばドライブ#3のアドレス(2, 2)のデータを更新する場合、まず、更新される前のドライブ#3の(2, 2)のデータとパリティを格納してあるドライブの(2, 2)のパリティを読みだし(1)、これらと更新する新しいデータとで排他的論理和をとり、新たなパリティを作成する(2)。パリティの作成完了後、更新する新しいデータをドライブ#3の(2, 2)に、新パリティをパリティを格納するドライブの(2, 2)に格納する(3)。

【0012】図13に示すように、このようなレベル5のディスクアレイでは、データの格納されているドライブ、パリティの格納されているドライブから古いデータ

とパリティを読みだすため、ディスクを平均1/2回転待ち、それから読みだしてパリティを作成する。この新しく作成したパリティを書き込むため更に一回転必要となり、データを書き替える場合最低で1.5回転待たなければならない。ドライブにおいては1.5回転ディスクの回転を待つということは非常に大きなオーバーヘッドとなる。このような書き込み時のオーバーヘッドを削減するため、書き込み先のアドレスを動的に変換する方法が考えられ、WO 91/20076に開示されている。

【0013】また、前記特開平4-230512においても、書き込み時において書き込みデータをそのまま、書き込みデータが書き込まれるアドレスではなく別のアドレスに、書き込むことにより書き込みオーバーヘッドを削減する方法について開示されている。CPU側から書き込むデータが送られてくるとすぐにパリティの更新を行ない、更新後のパリティを書き込む。このように、レベル5のディスクアレイでは、読みだしと比較し書き込み時ではパリティ生成とこの生成したパリティを書き込む処理のオーバーヘッドが非常に大きいため、CPUからの読みだし、書き込み要求が多いときには、この処理オーバーヘッドが性能低下の大きな原因となる。

【0014】本発明の目的は、レベル5のディスクアレイ装置において、書き込み時における処理のオーバーヘッドを減少させて、ディスクアレイ装置の性能向上を図ることにある。

【0015】本発明の他の目的は、障害ドライブ内のデータ復元用スペアドライブを装置の性能向上に利用することによってドライブ資源の有効活用を図ることにある。

【0016】

【課題を解決するための手段】本発明ではパリティグループを構成するドライブと二重化領域(スペース領域)のドライブにより論理グループを構成し、このスペース領域を有効に活用することにより、高信頼性を保ちながら、しかも、書き込み時のパリティ更新の開始時間を遅らせ、後のCPUからの読みだしまたは書き込み要求が少ないときにパリティ生成を行う。

【0017】具体的には、書き込み時に、論理グループ10を構成するSCSIドライブ12の中で、書き込むデータ(新データ)をとりあえずスペース領域に二重化して格納する。CPU1に対してはこの時点で書き込み処理を完了したと報告する。

【0018】また、パリティの作成および当該SCSIドライブ12へのパリティの書き込みは、新データのSCSIドライブへの書き込みとは独立のタイミングで、処理する。具体的には、ADC2のMP120が当該論理グループ10に対するCPU1からの読みだし/書き込み要求をカウントし、予めユーザまたはシステム管理者が設定した数より少ない場合で、しかも当該SCSIドライブ12に対し読みだしまたは書き込み要求が発

10

20

30

40

50

行されていないときにパリティの作成を行い、パリティの作成完了後当該SCSIドライブ12に対しパリティを書き込む。

【0019】パリティの書き込みの他の方法として、一定時間毎の割込み処理で行なってもよい。一日の中でCPUからの読みだしまたは書き込み要求数の少ない時間帯、あるいは一月の中で少ない日を予測し、スケジュール化しておけばよい。

【0020】パリティの作成およびそのパリティの当該SCSIドライブ12へのパリティの書き込みが完了する前に、当該論理グループ10において任意の1台のSCSIドライブに対し障害が発生し、その内部のデータが読み出せなくなった場合は、二重化データ以外のデータが格納されているSCSIドライブ12に対しては、前のパリティと、残っているデータから障害が発生したSCSIドライブ12内のデータを回復することが可能で、二重化されている新しいデータにおいては、障害が発生していない方のSCSIドライブ12内の新データにより回復することが可能である。

【0021】

【作用】本発明では、上記のようにデータの書き込みとパリティの作成およびSCSIドライブ12への書き込みを独立させることにより、ユーザ（CPU）からは書き込み時のパリティ作成によるオーバーヘッドはなくなる。これは、CPUからの読みだしまたは書き込み要求数には時間的変動があるため、読みだしまたは書き込み要求数が多いときには、書き込み処理におけるパリティの更新をその都度行わず、データの書き込みが完了した時点でCPUには終了を報告し、比較的読みだしまたは書き込み要求の数が少ないときまでパリティの更新を遅らせる。このパリティの更新はCPU側の関知はなく、ディスクアレイコントローラ2が独自に行う。

【0022】このため、CPU側から見たとき、従来のディスクアレイ（RAID方式）では図12に示すように書き込み時に平均1.5回転の回転待ち時間を必要としたのが、本発明によれば平均0.5回転の回転待ち時間ですむ。また、信頼性の面から見ても従来のディスクアレイ（RAID方式）と比較し、同等に向上させることが可能となる。

【0023】

【実施例】

〈実施例1〉以下本発明の一実施例を、図1～図5及び図13により説明する。図1において、本実施例はCPU1、ディスクアレイコントローラ（以下ADC）2、ディスクアレイユニット（以下ADU）3により構成される。ADU3は複数の論理グループ10により構成され、個々の論理グループ10はm台のSCSIドライブ12と、各々のSCSIドライブ12とADC2を接続するドライブパス9-1から4により構成される。なお、このSCSIドライブ12の数は本発明の効果を

るには、特に制限は無い。この論理グループ10は障害回復単位で、この論理グループ10内の各SCSIドライブ12内の各データによりパリティを作成する。本実施例ではm-1台の個々のSCSIドライブ12内のデータから各々のパリティが作成される。

【0024】次にADC2の内部構造について図1を用いて説明する。ADC2はチャンネルバスディレクタ5と2個のクラスタ13とバッテリバックアップ等により不揮発化された半導体メモリであるキャッシュメモリ7により構成される。このキャッシュメモリ7にはデータとアドレス変換用テーブルが格納されている。このキャッシュメモリ7およびその中のアドレス変換用テーブルはADC2内の全てのクラスタにおいて共有で使用される。クラスタ13はADC2内において独立に動作可能なパスの集合で、各クラスタ13間においては電源、回路は全く独立となっている。クラスタ13はチャンネル、キャッシュメモリ7間のパスである、チャンネルパス6と、キャッシュメモリ7、SCSIドライブ12間のパスであるドライブパス6-1から4が、それぞれ、2個ずつで構成されている。それぞれのチャンネルパス6-1から4とドライブパス8はキャッシュメモリ7を介して接続されている。CPU1より発行されたコマンドは外部インターフェースパス4を通してADC2のチャンネルバスディレクタ5に発行される。ADC2は2個のクラスタ13により構成され、それぞれのクラスタは2個のパスで構成されるため、ADC2は合計4個のパスにより構成される。このことから、ADC2ではCPU1からのコマンドを同時に4個まで受け付けることが可能である。そこで、CPU1からコマンドが発行された場合ADC2内のチャンネルバスディレクタ5によりコマンドの受付が可能かどうか判断する。

【0025】図2は図1のチャンネルバスディレクタ5と1クラスタ13-1内の内部構造を示した図である。図2に示すように、CPU1からADC2に送られてきたコマンドはインターフェースアダプタ（以下IF Adp）15により取り込まれ、マイクロプロセッサであるMP120はクラスタ内の外部インターフェースパス4の中で使用可能なパスがあるかを調べ、使用可能な外部インターフェースパス4がある場合はMP120はチャンネルバススイッチ16を切り換えてコマンドの受け付け処理を行ない、受け付けられない場合は受付不可の応答をCPU1へ送る。

【0026】本実施例ではADU3を構成するSCSIドライブ12はSCSIインターフェースのドライブを使用する。CPU1をIBMシステム9000シリーズのような大型汎用計算機とした場合、CPU1からはIBMオペレーティングシステム（OS）で動作可能なチャンネルインターフェースのコマンド体系にのっとってコマンドが発行される。そこで、SCSIドライブ12をSCSIインターフェースのドライブを使用した場合、

CPU1からのコマンドを、SCSIインターフェースのコマンド体系にのっとったコマンドに変換する必要が生じる。この変換はコマンドのプロトコル変換と、アドレス変換に大きく分けられる。以下にアドレス変換について説明する。

【0027】CPU1から指定されるアドレスは、図12に示すように当該データが格納されているトラックが所属するシリンダの位置と、そのシリンダ内において当該データが格納されているトラックを決定するヘッドアドレスと、そのトラック内のレコードの位置を特定する。具体的には要求データが格納されている当該ドライブの番号(CPU指定ドライブ番号)と当該ドライブ内のシリンダ番号であるシリンダアドレス(CC)とシリンダにおいてトラックを選択するヘッドの番号であるヘッドアドレス(HH)とレコードアドレス(R)からなるCCHHRである。

【0028】従来のCKDフォーマット対応の磁気ディスクサブシステム(IBM3990-3390)ではこのアドレスに従ってドライブへアクセスすれば良い。しかし、本実施例では複数のSCSIドライブ12により従来のCKDフォーマット対応の磁気ディスクサブシステムを論理的にエミュレートする。つまり、ADC2は複数のSCSIドライブ12が、従来のCKDフォーマット対応の磁気ディスクサブシステムで使用されているドライブ1台に相当するようにCPU1にみせかける。このため、CPU1から指定してきたアドレス(CCHHR)をSCSIドライブのアドレスにMP1 20は変換する。このアドレス変換には以下に示すようなアドレス変換用のテーブル40(以下アドレステーブルとする)が使用される。

【0029】ADC2内のキャッシュメモリ7には、その内部の適当な領域に図3に示すようなアドレステーブル40が格納されている。本実施例では、CPU1が指定してくるドライブはCKDフォーマット対応の単体ドライブである。しかし、本発明ではCPU1は単体と認識しているドライブが、実際は複数のSCSIドライブ12により構成されるため、論理的なドライブとして定義される。このため、ADC2のMP1 20はCPU1より指定してきたアドレス(CPU指定ドライブ番号41とCCHHR46)をSCSIドライブ12に対するSCSIドライブアドレス42(SCSIドライブ番号43とそのSCSIドライブ内のアドレス(以下SCSI内Addrとする)44)に変換する。

【0030】アドレステーブル40はCPU1が指定するCPU指定ドライブ番号41とSCSIドライブアドレス42により構成される。SCSIドライブアドレス42はSCSIドライブ12のアドレスであるSCSIドライブ番号43とそのSCSIドライブ内の実際にデータが格納されているアドレスである、SCSI内Addr44と、論理グループ10内において、そのSCS

I内Addr44により決定されるパリティグループにおけるパリティが格納されているSCSIドライブ番号(パリティドライブ番号50)と、二重化領域(スペース領域)が格納されているSCSIドライブ番号(スペースドライブ番号51)により構成されている。このアドレステーブル40では、論理アドレス45によりSCSIドライブ番号43とSCSI内Addr44を決定する。アドレステーブル40のSCSIドライブアドレス42に登録されているSCSIドライブ番号43のSCSIドライブ12により論理グループ10は構成される。

【0031】この論理グループ10内の同一SCSI内Addr44において、パリティが格納されているSCSIドライブ番号43をパリティドライブ番号50に登録し、スペース領域が確保されているSCSIドライブ番号43をスペースドライブ番号51に登録する。スペースドライブ番号51にはスペース領域が確保されているSCSIドライブ番号の他にSDフラグ53により構成される。SDフラグ53は、スペース領域に格納されているデータが有効で、書き込み処理においてスペース領域として使用できない場合はオン(1)となり、無効で、スペース領域として使用可能な場合はオフ(0)となる。この論理グループ10はデータとこのデータと関連するパリティにより構成されるパリティグループとスペース領域により構成される。

【0032】論理アドレス45にはCPU1から指定されるアドレスである、CPU指定ドライブ番号41とCCHHRの中でCCHHR46が登録されており、それ以外にはこの論理アドレス45のデータがキャッシュメモリ7内に存在する場合の、そのデータのキャッシュメモリ7内のアドレスを格納するキャッシュアドレス47と、キャッシュメモリ7内にその論理アドレス45のデータを保持している場合オン(1)が登録されるキャッシュフラグ48と、その論理アドレス45にはスペース領域が確保されている場合オン(1)となる無効フラグ49とキャッシュメモリ7内の書き込みデータがドライブに書き込まれている場合オン(1)となるドライブフラグ52により構成される。

【0033】以上のようにアドレステーブル40によりCPU指定ドライブ番号41とCCHHR46を論理アドレス45に変換し、そのデータが実際に格納されているSCSIドライブ番号43とSCSI内Addr44を決定する。

【0034】例えば、図3においてCPU1からCPU指定ドライブ番号41としてDrive#1、CCHHR46がADR8のデータに対し要求を発行してきた場合、アドレステーブル40においてCPU指定ドライブ番号41がDrive#1の領域の各論理アドレス45内のCCHHR46を調べ、CCHHR46がADR8の論理アドレス45を探す。図3においては論理アドレ

ス45としてData#23(D#23)がCCHHR46がADR8となっており、Data#23(D#23)が当該論理アドレス45である。

【0035】このData#23(D#23)はアドレステーブル40からSD#2のSCSIインターフェースのSCSIドライブ12内のSCSI内Addr44としてDADR8に該当することが分かり、物理的なアドレスへ変換される。また、このData#23(D#23)に対応するパリティは、パリティドライブ番号50からData#23(D#23)と同一のSCSI内Addr44のDADR8のSD#4のSCSIドライブ12に格納されており、スペアドライブ番号51から、SDフラグ53がオン(1)のため、SD#4、5のSCSI内Addr44がDADR8に二重化して格納されているデータは有効で、この領域は二重化領域(スペア領域)として使用することは禁止されている。

【0036】このように、CPU1から指定されたアドレスを論理アドレス45に変換し、実際に読みだし/書き込みを行うSCSIドライブ12の物理的なアドレスに変換した後、SD#2のSCSIドライブ12のData#23(D#23)に対し読み出しまたは書き込み要求が発行される。この時アドレステーブル40においてData#23(D#23)の論理アドレス45ではキャッシュフラグ48がオン(1)のためこのデータはキャッシュメモリ7内のCADR2、1に存在する。もし、キャッシュフラグ48がオフ(0)であればCADR2、4のキャッシュメモリ7内には、当該データは存在しない。また、このデータは無効フラグ49がオフ(0)のため、このデータは有効となる。さらにドライブフラグ52がオン(1)のため、このデータはキャッシュメモリ7からドライブに既に書き込まれている。

【0037】この、アドレステーブル40はシステムの電源をオンした時に、MP1 20により論理グループ10内のある特定のSCSIドライブ12から、キャッシュメモリ7にCPU1は関知せず自動的に読み込まれる。一方、電源をオフする時はMP1 20によりキャッシュメモリ7内のアドレステーブル40を、読み込んだきたSCSIドライブ12内の所定の場所にCPU1は関知せずに自動的に格納する。

【0038】次に、ADC2内での具体的なI/O処理について図1、図2を用いて説明する。CPU1より発行されたコマンドはIFAdp15を介してADC2に取り込まれ、MP1 20により読み出し要求が書き込み要求が解釈される。まず、読み出し要求の場合の処理方法を以下に示す。

【0039】MP1 20が読み出し要求のコマンドを認識すると、MP1 20はCPU1から送られてきたCPU指定ドライブ番号41とCCHHR46(以下両方を併せてCPU指定アドレスとする)についてアドレステーブル40を参照し、当該データの論理アドレス4

5への変換を行ない、論理アドレス45に登録されているキャッシュメモリ7内に存在するかどうかキャッシュフラグ48を調べ、判定する。

【0040】キャッシュフラグ48がオンでキャッシュメモリ7内に格納されている場合(キャッシュヒット)は、MP1 20がキャッシュメモリ7から当該データを読みだす制御を開始し、キャッシュメモリ7内に無い場合(キャッシュミス)は当該ドライブ12へその内部の当該データを読みだす制御を開始する。

【0041】キャッシュヒット時、MP1 20はアドレステーブル40によりCPU1から指定してきたCPU指定アドレスから論理アドレス45に変換し、論理アドレス45内のキャッシュアドレス47によりキャッシュメモリ7のアドレスに変換し、キャッシュメモリ7へ当該データを読み出しに行く。具体的にはMP1 20の指示の元でキャッシュアダプタ回路(CAdp)24によりキャッシュメモリ7から当該データは読み出される。

【0042】CAdp24はキャッシュメモリ7に対するデータの読みだし、書き込みをMP1 20の指示で行う回路で、キャッシュメモリ7の状態の監視、各読みだし、書き込み要求に対し排他制御を行う回路である。CAdp24により読み出されたデータはデータ制御回路(DCC)22の制御によりチャネルインターフェース回路(CHIF)21に転送される。CHIF21ではCPU1におけるチャネルインターフェースのプロトコルに変換し、チャネルインターフェースに対応する速度に速度調整する。具体的にはCPU1、ADC2間のチャネルインターフェースを光のインターフェースにした場合、光のインターフェースのプロトコルをADC2内では電気処理でのプロトコルに変換する。CHIF21におけるプロトコル変換および速度調整後は、チャネルバスディレクタ5において、チャネルバススイッチ16が外部インターフェースバス4を選択しIFAdp15によりCPU1へデータ転送を行なう。

【0043】一方、キャッシュミス時はキャッシュヒット時と同様にアドレステーブル40により、CPU指定アドレスを論理アドレス45に変換し、論理アドレス45から当該SCSIドライブ番号とそのSCSIドライブ内の実際にデータが格納されているSCSI内Addr44を認識し、そのアドレスに対し、MP1 20はDriveIF28に対し、当該ドライブ12への読み出し要求を発行するように指示する。DriveIF28ではSCSIの読み出し処理手順に従って、読み出しコマンドをドライブユニットバス9-1または9-2を介して発行する。DriveIF28から読み出しコマンドを発行された当該SCSIドライブ12においては指示されたSCSI内Addr44へシーク、回転待ちのアクセス処理を行なう。当該SCSIドライブ12におけるアクセス処理が完了した後、当該SCSI

ドライブ12は当該データを読み出しドライブユニットバス9を介してDrive IF28へ転送する。

【0044】Drive IF28では転送されてきた当該データをSCSIドライブ側のキャッシュアダプタ回路(C Adp)14に転送し、(C Adp)14ではキャッシュメモリ7にデータを格納する。この時、C Adp14はキャッシュメモリ7にデータを格納することをMP1 20に報告し、MP1 20はこの報告を元にアドレステーブル40のCPUが読みだし要求を発行したCPU指定アドレスに対応した論理アドレス45のキャッシュフラグ48をオン(1)にし、キャッシュアドレス47にキャッシュメモリ7内のデータを格納したアドレスを登録する。キャッシュメモリ7にデータを格納し、アドレステーブル40のキャッシュフラグ48をオン(1)にし、キャッシュメモリ7内のアドレスを更新した後はキャッシュヒット時と同様な手順でCPU1へ当該データを転送する。

【0045】一方書き込み時は以下のように処理される。書き込み処理はユーザが書き込み先のアドレス(CPU指定アドレス)を指定し、その位置にユーザはデータを書き込んでいると認識する。つまりユーザは固定の位置にアドレスを指定していると認識している。

【0046】CPU1から、指定アドレスすなわち、図3に示すアドレステーブル40においてCPU指定ドライブ番号41がドライブ#1でCCHHR46がADR8に対して、書き込み命令が発行されたとする。まず、ADC2のMP1 20はCPU1からドライブ#1のCCHHR46がADR8に対し書き込み要求のコマンドを受け取った後、コマンドを受け取ったMP1 20が所属するクラス13内の各チャンネルバス6において処理可能かどうかを調べ、可能な場合は処理可能だという応答をCPU1へ返す。CPU1では処理可能だという応答を受け取った後にADC2へデータを転送する。この時、ADC2ではMP1 20の指示によりチャンネルバスディレクタ5において、チャンネルバススイッチ16が当該外部インターフェースバス4とIF Adp15を当該チャンネルバス6と接続しCPU1とADC2間の接続を確立する。

【0047】CPU1とADC2間の接続を確立後CPU1からのデータ転送を受け付ける。CPU1から転送されてきた書き込みデータ(以下新データとする)はMP1 20の指示により、CH IF21によりプロトコル変換を行ない、外部インターフェースバス4での転送速度からADC2内での処理速度に速度調整する。CH IF21におけるプロトコル変換および速度制御の完了後、データはDCC22によるデータ転送制御を受け、C Adp24に転送され、C Adp24によりキャッシュメモリ7内に格納される。

【0048】この時、CPU1から送られてきた情報が、CPU指定アドレスの場合は、データが転送されて

くる前に必ずCPU指定アドレスがCPU1より転送されているため、読みだしと同様にアドレステーブル40によりアドレス変換を行い、論理アドレス45に変換する。また、CPU1から送られてきた情報がデータの場合は、キャッシュメモリ7に格納したアドレスを上記アドレス変換により変換した論理アドレス45のキャッシュアドレス47に登録する。この時、書き込みデータをキャッシュメモリ7内に保持するときは、論理アドレス45のキャッシュフラグ48をオン(1)とし、保持しない場合はキャッシュフラグ48をオフ(0)とする。

【0049】なお、キャッシュメモリ7内に保持されている新データに対し、さらに書き込み要求がCPU1から発行された場合は、キャッシュメモリ7内に保持されている新データを書き替える。

【0050】キャッシュメモリ7に格納された新データは、この新データにより新しくパリティを更新し(以下更新されたパリティを新パリティとする)、以下のように論理グループ10内のSCSIドライブ12へ新データと新パリティを格納する。

【0051】図3に示すようにスペース領域及びパリティはデータと同じ様に扱われ、論理グループ10内の各SCSIドライブ12に分散して格納されている。論理グループ10を構成する各SCSIドライブ12内のデータについては行方向(同一SCSI内Addr44)にパリティグループが構成され、このパリティグループ内のデータにおいてパリティが作成される。つまり、パリティグループはデータとパリティにより構成され、論理グループ10はパリティグループとスペース領域で構成される。

【0052】図3において具体的な例を示す。SCSI内Addr44がDADR1については、SD#1のSCSIドライブ12に格納されているData#1(D#1)と、SD#2のSCSIドライブ12に格納されているData#2(D#2)と、SD#3のSCSIドライブ12に格納されているData#3(D#3)によりパリティが作成される。このパリティがSD#6のSCSIドライブ12に格納され、これらがパリティグループを構成する。また、論理グループ10は、SD#4のSCSIドライブ12に確保されているスペース領域(S)と、SD#5のSCSIドライブ12に確保されているスペース領域(S)と、上記パリティグループにより構成される。

【0053】MP1 20はアドレステーブル40を参照し、データ、スペース領域、パリティが格納されているSCSIドライブ12を認識する。具体的には、アドレステーブル40においてCPU指定アドレスが指定したCPU指定ドライブ番号41に対応する領域において、CPU指定アドレスで指定したCCHHR46と一致しているSCSIドライブアドレス42に登録されている論理アドレス45をMP1 20は探す。MP1

20はCPU指定アドレスから論理アドレス45への変換後、論理アドレス45からその論理アドレス45が格納されているSCSIドライブ番号43と、そのSCSIドライブ12内の物理的なアドレスであるSCSI内Addr44に変換する。

【0054】パリティは、論理グループ10を構成する各SCSIドライブ12において、各SCSI内Addr44が同一のデータに対し作成され、そのパリティもデータと同一のSCSI内Addr44に格納される。このため、アドレステーブル40のパリティドライブ番号50とスペースドライブ番号51において、パリティおよびスペース領域はそれぞれが格納されているSCSIドライブ番号43のみが登録されている。そこで、MP120ではアドレステーブル40によりパリティドライブ番号50とスペースドライブ番号51を決定することが可能となる。つまり、パリティドライブ番号50とスペースドライブ番号51を決定することにより、パリティおよびスペース領域はデータと同一のSCSI内Addr44に格納されていることから、それらのアドレスを決定することになる。このように、データ、スペース領域、パリティが格納されているSCSIドライブ12を認識した後は、DriveIF28に対し、各々の当該SCSIドライブ12に対し書き込み処理を行なうように指示する。

【0055】本発明における書き込み処理とは、論理グループ10において新データをキャッシュメモリ7から実際にSCSIドライブ12に書き込む処理と、新データの書き込みによりパリティを新たに作り直さなければならないため、この新パリティを作成するための書き込み前のデータ（以下旧データ）と書き込み前のパリティ（以下旧パリティとする）を読み出し、新パリティを作成する処理と、書き込み後の新パリティを実際にSCSIドライブ12に書き込む一連の処理をいう。この新データをキャッシュメモリ7に格納した後の一連の書き込み処理のフローチャートを図5に示す。

【0056】図4に示すようにCPU1からSD#1のSCSIドライブ12のData#1（D#1）の論理アドレスに対し、ND#1のデータの書き込み要求が発行された場合、先に示したように、CPU1からキャッシュメモリ7に新データは一旦格納される。新データのキャッシュメモリ7への格納後、書き込み処理は次の手順で行っていく。キャッシュメモリ7に新データ（ND#1）が格納された後、MP120は書き込み処理に入り、アドレステーブル40によりSD#1のSCSIドライブ12が所属する論理グループ10で、Data#1（D#1）のSCSI内Addr44であるDADR1に対し、スペース領域が確保されているSD#4、5のSCSIドライブ12に対し、使用権の獲得を行なう。

【0057】SD#4、5のSCSIドライブ12の使

用権を獲得した後は図5のフローチャートに示すように書き込み処理500を行なっていく。まず、MP120はアドレステーブル40のスペースドライブ番号51のSDフラグ53をチェックし、SDフラグ53がオフ（0）の場合、スペース領域として使用でき、オン

（1）の場合は使用できないと判断する（502）。MP120は、このSDフラグ53によりSD#4、5のSCSIドライブ12にスペース領域が確保されているかを判断する。そして、SDフラグ53がオフ（0）の場合、キャッシュメモリ7に格納されているND#1をこのSD#4、5に二重化して書き込み（504）、MP120はCPU1へ書き込み完了の報告を行う（508）。

【0058】一方、ステップ502のチェックにより、SDフラグ53がオン（1）の場合は、図13に示すように、書き込む新データ（ND#1）をキャッシュメモリ7に書き込んだ後、MP120は優先的に前の書き込み処理におけるパリティの作成を指示し、このパリティを当該SCSIドライブ12に書き込む（1310）。前の書き込み処理におけるパリティの作成が完了し、当該SCSIドライブへの書き込みが完了するとMP120は、アドレステーブル40においてスペースドライブ番号51にSDフラグ53をオフ（0）とし（1308）、キャッシュメモリ7に格納されているND#1を二重化して書き込み（1316）、MP120はCPU1へ書き込み完了の報告を行う（1320）。

【0059】そこで、次にSCSIドライブ12への新データ（ND#1）の書き込み方法を示す。MP120はアドレステーブル40のSDフラグ53がオフ（0）になっているのを確認後、DriveIF28に書き込むデータである新データ（ND#1）をスペース領域が確保されているSD#4とSD#5のSCSIドライブ12に書き込むよう指示する。DriveIF28ではSCSIの書き込み手順にしたがってドライブユニットパス9-1から4の中の2本を介してSD#4とSD#5のSCSIドライブ12に書き込みコマンドを発行する。

【0060】DriveIF28から書き込みコマンドを発行された当該SCSIドライブ12では、DriveIF28から送られてきたCPU1が書き込み先のアドレスとして指定したCPU指定アドレスを論理アドレス45に変換し、論理アドレス45であるData#1に対応するSCSI内Addr44のDADR1へシーク、回転待ちのアクセス処理を行なう。SD#4、5のSCSIドライブ12においてアクセス処理が完了し書き込みが可能になり次第、CAdp14はキャッシュメモリ7から書き込む新データ（ND#1）を読み出してDriveIF28へ転送し、DriveIF28では転送されてきた新データ（ND#1）をドライブユ

ニットパス9-1から4の中の2本を介してSD#4, 5のSCSIドライブ12へ転送する。新データ(ND#1)のSD#4, 5のSCSIドライブ12への書き込みが完了すると、SD#4, 5のSCSIドライブ12はDrive IF28に完了報告を行ない、Drive IF28がこの完了報告を受け取ったことを、MP120に報告する。

【0061】この時、アドレステーブル40において、書き込み前の旧データの論理アドレス45(Data#1, D#1)の無効フラグをオン(1)とし、書き込み前の論理アドレス45(Data#1, D#1)内のCCHHR46のアドレスを、新データ(ND#1)を二重化して書き込んだスペース領域の2つの論理アドレス45のCCHHR46に登録し、無効フラグをオフ(0)とし、ドライブフラグ52をオン(1)とする。また、キャッシュメモリ7内に書き込む新データ(ND#1)を保持する場合は、書き込み後の2つの論理アドレス45内のキャッシュアドレス47に、キャッシュメモリ7内の新データ(ND#1)が格納されているアドレスに登録し、キャッシュフラグ48をオン(1)とする。この新データ(ND#1)をキャッシュメモリ7上に残さない場合、MP120はこの報告を元にアドレステーブル40のキャッシュフラグ48をオフ(0)にする。さらに、書き込みを行った論理グループ10内のSCSI内Addr44に対し、スペースドライブ番号51のSDフラグをオン(1)とする。

【0062】上記のようにアドレステーブル40の更新が完了し、MP120はSD#4とSD#5の両方のSCSIドライブ12からの完了報告を受け取った後、CPU1に対し擬似的に書き込み処理の終了報告を行う。SD#4, 5のSCSIドライブ12に対する新データ(ND#1)の格納は完了しても、キャッシュメモリ7内には新データ(ND#1)が存在しており、パリティの更新はキャッシュメモリ7内に格納されている新データ(ND#1)で行なう。

【0063】以上のようにMP120がCPU1に対し擬似的に終了報告を行った後は、CPU1は書き込み処理を終了したと認識しているが、MP120は新しいパリティを作成して、当該SCSIドライブ12書き込んでいないため、まだ書き込み処理は終了していない。そこで、MP120がCPU1に対し終了報告を行った後に、MP120が独自に新しいパリティを作成し、当該SCSIドライブ12に書き込む。この方法について以下に説明する。

【0064】MP120はCPU1に対し書き込み処理の終了報告を行った後、図5に示すようにCPU1からの読みだし、書き込み要求の状況(I/O状況)を監視する(510)。具体的には、MP120が当該論理グループ10に対するCPU1からの読みだし、書き込み要求の単位時間当りの回数を数える。当該論理グルー

プ10に対するこの回数が予めユーザまたはシステム管理者により設定された数より少なく、パリティの作成およびこの作成したパリティを書き込むSCSIドライブ12が所属する論理グループ10に対し、読みだし、書き込み要求がCPU1より発行されていない場合、パリティの作成およびパリティの当該SCSIドライブ12への書き込み処理を開始する。

【0065】新しくパリティを作成する場合、CPU1から書き込み先に指定されたアドレスに書き込まれているデータ(旧データ)と、更新するパリティ(旧パリティ)を読みだし、新しく作成したパリティ(新パリティ)をSCSIドライブ12に書き込む。この時、旧データと旧パリティを読みだすSCSIドライブ12と新パリティを書き込む当該SCSIドライブ12に対し、MP120はCPU1からの読みだし、書き込み要求と同様な擬似的な読みだし、書き込み要求を発行する。この、擬似的な読みだし、書き込み要求が発行されているSCSIドライブ12に対しCPU1から読みだしまたは書き込み要求が発行された場合は、MP120はCPU1からの読みだし、書き込み要求を受付、処理の待ち行列とする。

【0066】次に新パリティの作成および当該SCSIドライブ12への作成したパリティの書き込み方法を具体的に示す。MP120はSD#1のSCSIドライブ12に対して旧データの読み出しと、SD#6のSCSIドライブ#12に対して旧パリティの読み出し要求を発行するようにDrive IF28に指示する(514)。

【0067】Drive IF28から読み出しコマンドを発行された当該SCSIドライブ12ではDrive IF28から送られてきたSCSI内Addr44ヘシーク、回転待ちのアクセス処理を行なう。キャッシュメモリ7内には新データ(ND#1)が存在しており、パリティの更新はキャッシュメモリ7内に格納されている新データ(ND#1)で行なう。

【0068】もし、キャッシュメモリ7に新データ(ND#1)が存在していない場合はスペース領域に二重化して書き込まれているデータをキャッシュメモリ7に読みだす。

【0069】SD#1, 6のSCSIドライブ12においてシーク、回転待ちのアクセス処理が完了し旧データ(D#1)および旧パリティ(P#1)の読み出しが可能になり次第、旧データ(D#1)および旧パリティ(P#1)を読み出し、キャッシュメモリ7に格納する。SD#1のSCSIドライブ12から旧データ(D#1)を、SD#6のSCSIドライブ12から旧パリティ(P#1)を読み出し、それぞれをキャッシュメモリ7に格納した後、MP120はキャッシュメモリ7内に格納されている書き込む新データ(ND#1)とで排他的論理和により、更新後の新パリティ(NP#1)

を作成するようにPG36に指示を出し、PG36において新パリティ(NP#1)を作成しキャッシュメモリ7に格納する(516)。

【0070】新パリティ(NP#1)をキャッシュメモリ7に格納した後、MP1 20では新パリティ(NP#1)を格納する論理アドレス45のキャッシュアドレス47にキャッシュメモリ7内の新パリティ(NP#1)を格納したアドレスを登録し、キャッシュフラグ48をオン(1)とし、無効フラグ49とドライブフラグ52をオフ(0)とする(518)。新パリティ(NP#1)の作成が完了したことを認識し、SD#6のSCSIドライブ12にIO要求が発行されていない時に、更新後の新パリティ(NP#1)を書き込むようにDrive IF28に対し指示する。

【0071】SD#6のSCSIドライブ12への更新後の新パリティ(NP#1)の書き込み方法(520)は、先に述べた書き込む新データ(ND#1)をSD#4, 5のSCSIドライブ12に書き込んだ方法と同じである。新パリティ(NP#1)の作成が完了したらMP1 20はDrive IF28にSD#6のSCSIドライブ12に対し、書き込みコマンドを発行するように指示し、当該SCSIドライブ12では指示SCSI内Addr44へシーク、回転待ちのアクセス処理を行う。新パリティ(NP#1)は既に作成されキャッシュメモリ7に格納されており、SD#6のSCSIドライブ12におけるアクセス処理が完了した場合、C Adp14はキャッシュメモリ7から新パリティ(NP#1)を読み出してDrive IF28へ転送する。Drive IF28では転送されてきた新パリティ(NP#1)をドライブユニットパス9-1から4の中の1本を介してSD#6のSCSIドライブ12へ転送する(522)。

【0072】新パリティ(NP#1)のSD#6のSCSIドライブ12への書き込みが完了すると、SD#6のSCSIドライブ12はDrive IF28に完了報告を行ない、Drive IF28がこの完了報告を受け取ったことを、MP1 20に報告する。この時、MP1 20は、この新データ(ND#1)をキャッシュメモリ7上に残さない場合は、この報告を元にアドレステーブル40のキャッシュフラグ48をオフ(0)にし、キャッシュメモリ7上に残す場合はオン(1)とする。さらに、アドレステーブル40において、書き込み後の新パリティ(NP#1)の論理アドレス45の無効フラグをオフ(0)とし、ドライブフラグ52をオン(1)とする(524)。

【0073】このような、新しく作成した新パリティ(NP#1)を当該SCSIドライブ12に書き込んだ後は、旧データ(D#1)が格納されていたSD#1のSCSIドライブ12の旧データ(D#1)と二重化されている新データ(ND#1)が格納されているSD#

4, 5のSCSIドライブ12の中でSCSIドライブ番号の小さいSD#4のSCSIドライブ12に格納されている新データ(ND#1)をスペース領域として解放し、次の書き込み処理用にスペース領域として登録する。この登録の方法は、MP1 20がアドレステーブル40においてSD#1とSD#4のSCSI内Addr44がD ADR1の旧データ(D#1)と、二重化されていた新データ(ND#1)の一方が格納されている論理アドレス45において、無効フラグをオン(1)とし、さらにスペースドライブ番号51にSD#1とSD#4を登録し、SDフラグをオフ(0)にする(526)。

【0074】以上のように書き込み処理時に書き込む新データ(ND#1)を一旦二重化して論理グループ10内に格納しておき、後で比較的CPU1からの読みだし、書き込み要求が少ないときに、新パリティ(NP#1)を作成し、SCSIドライブ12に格納することで、書き込み処理時の応答時間が短縮され、従来のように新パリティ(NP#1)の作成により他の読みだし、書き込み処理がまたされることが減少する。

【0075】もし、新パリティ(NP#1)を当該SCSIドライブ12への書き込み前に論理グループ10内のSCSIドライブ12に障害が発生した場合は図14に示すように回復処理を行う。図4(a)に示すようにSD#6のSCSIドライブ12に対し新パリティ(NP#1)を書き込む前に(1402)、SD#1, 2, 3の内のどれか1台のSCSIドライブ12に障害が発生した場合(1406)、障害が発生していないSCSIドライブ12からのデータと、旧パリティから障害が発生したSCSIドライブ12内のデータは回復することが可能となる(1410)。例えばSD#1のSCSIドライブ12に障害が発生した場合、SD#2, 3のD#2, D#3とSD#6のSCSIドライブ1からの旧パリティ(P#1)からSD#1のSCSIドライブ12内のデータであるD#1を回復することが可能となる。また、新データ(ND#1)が二重化して格納されているSD#4, 5のどちらか一方に障害が発生した場合は、二重化データの一方のデータにより回復することが可能となる(1412)。

【0076】図4(b)に示すようにSD#6のSCSIドライブ12に対し新パリティ(NP#1)を書き込んだ後はSD#2, 3, 5のSCSIドライブ12内のデータ(D#2, 3, ND#1)に対し新パリティ(NP#1)が作成されてSD#6のSCSIドライブ12に格納されており、SD#2, 3, 5のSCSIドライブ12の中の1台のSCSIドライブ12に障害が発生した場合、残りの正常なSCSIドライブ12内のデータとSD#6のSCSIドライブ12内のパリティにより、障害が発生したSCSIドライブ12内のデータは回復される。

【0077】例えばSD#2のSCSIドライブ12に障害が発生した場合、SD#2のSCSIドライブ12内のデータ(D#2)は、SD#3のSCSIドライブ12内のデータ(D#3)とSD#4のSCSIドライブ12内のデータ(ND#1)と、SD#6のSCSIドライブ12内のパリティ(NP#1)から障害が発生したSD#2のSCSIドライブ12内のデータ(D#2)は回復される。

【0078】本発明のように書き込み処理においてスペース領域に書き込みデータ(新データ)をとりあえず二重化して格納し、この段階でCPU1に対し書き込み処理の終了報告を行うことにより、CPU1にとってはこの二重化してSCSIドライブ12に書き込む時間が書き込み処理時間になる。従来のアレイディスクでは図12に示すように書き込み時に平均1.5回転の回転待ち時間が必要としたのが、もし、論理グループ10を構成するSCSIドライブ12の回転を同期させた場合は、回転待ちは平均0.5回転となる。また、新パリティをSCSIドライブ12に書き込む前に論理グループ10を構成するSCSIドライブ12に障害が発生しても、先に述べたように旧パリティと二重化された新データにより従来のアレイディスクと同様に、障害回復を行うことが可能となる。

【0079】本実施例ではパリティおよびパリティの作成に関与したデータとスペース領域は、論理グループ10を構成する各SCSIドライブ12において、各SCSI内Addr44を同一とした。しかし、アドレステーブル40の各論理アドレス45とパリティドライブ番号50、スペースドライブ番号51において、論理グループ10のアドレスを付加することにより、論理グループ10を構成する各SCSIドライブ12においてSCSI内Addr44を任意とすることが可能となる。

【0080】本実施例では書き込み時の回転待ち時間を減少させる目的で、書き込みデータを一旦二重化してSCSIドライブ12に書き込み、後の適当なタイミングでパリティを更新し、パリティの更新後は二重化した書き込みデータの一方を開放する方法について説明した。本発明は、上記のような性能向上の観点とは別に、以下に示すような応用例も考えられる。

【0081】パリティによる信頼性と、二重化による信頼性では、二重化による信頼性の方が信頼性向上のために必要とする容量は大きい、信頼性は高くなる。この特徴を活かし、本発明の応用として、最新に書き込まれたデータまたは、頻繁に書き込み要求が発行されるデータについては、二重化されているため、高信頼とし、あまり書き込み要求が発行されないデータについては、二重化ほど高信頼ではないが、信頼性向上のために必要とする容量が二重化と比較して少なくすむ、パリティにより信頼性を確保する。つまり、あまり書き込み要求が発行されないデータに対しては、信頼性は二重化ほど高

くないが、パリティを付けるだけですむ、パリティで信頼性を確保し、最新に書き込まれたデータまたは、頻繁に書き込み要求が発行されるデータについては、パリティと比較し信頼性向上のために容量を多く必要とするが、高信頼である二重化で高信頼とする、二段階の信頼性のレベルを設定することも可能である。

【0082】〈実施例2〉本発明の他の実施例を図6を中心にして説明する。本実施例では実施例1で示したシステムにおいて、SCSIドライブ12に障害が発生した時に、その障害が発生したSCSIドライブ12内のデータを回復し、それを格納するための領域にスペース領域を使用する例を示す。

【0083】本発明では図3に示すように、論理グループ10内の各SCSIドライブ12では、その内部の各々対応する同一SCSI内Addr44のデータによりパリティグループを構成する。具体的にはSD#1, 2, 3のSCSIドライブ12内のData#1, 2, 3(D#1, 2, 3)でPG36によりパリティ#1(P#1)が作られSD#6のSCSIドライブ12内に格納される。本実施例ではパリティは奇数パリティとし、Data#1, 2, 3(D#1, 2, 3)の各データにおける各々対応するビットについて1の数を数え、奇数であれば0、偶数であれば1とする(排他的論理和)。もし、SD#1のSCSIドライブ12に障害が発生したとする。この時、Data#1(D#1)は読み出せなくなる。

【0084】本実施例ではパリティグループ当りにパリティを1個しか持っていないため、1台のSCSIドライブ12の障害はデータを復元できるが、データの復元が完了する前に更にもう一台のSCSIドライブ12に障害が発生した場合復元出来ない。そこで、このような場合、2台目のSCSIドライブ12障害が発生する前に、残りのData#2, 3(D#2, 3)とパリティ#1(P#1)をキャッシュメモリ7に転送し、MP120はPG36に対しData#1(D#1)を復元する回復処理を早急に行なうように指示する。この回復処理を行ないData#1(D#1)を復元した後は、MP120はこのData#1(D#1)をSD#4または6のどちらかのスペース領域に格納する。

【0085】これにより、スペース領域を実施例1で示したような、書き込み処理時の回転待ち時間を短縮させるためだけではなく、SCSIドライブ12に障害が発生したときに、復元したデータを格納するためのスパー領域としても活用する。この様に、MP120がスペース領域に回復したData#1(D#1)を格納した後は、キャッシュメモリ7にある図3に示すアドレステーブル40において、スペースドライブ番号51の中で、回復データを格納した方のスペースドライブ番号51を削除し、この削除したドライブ番号に対する論理アドレス45に、回復したData#1(D#1)の論理

アドレス45の内容を複写する。

【0086】図6に示すようにSD#1のSCSIドライブ12にはData#1(D#1)の他にスペース領域、パリティ、Data#13、16、19、22(D#13、16、19、22)が格納されている。スペース領域については回復処理を行ない復元する必要はない。パリティ#3(P#3)はSD#3、4、5のSCSIドライブ12からData#7、8、9(D#7、8、9)を読み出して新たに作成しSD#2か6のSCSIドライブ12のスペース領域に格納する。Data#13(D#13)はSD#3、5、6のSCSIドライブ12からパリティ#5(P#5)、Data#14、15(D#14、15)を読み出して、回復処理を行ない復元し、SD#2または4のSCSIドライブ12のスペース領域に格納する。Data#16(D#16)はSD#2、4、6のSCSIドライブ12からData#17(D#17)、パリティ#6(P#6)、Data#18(D#18)を読み出して、回復処理を行ない復元し、SD#3または5のSCSIドライブ12のスペース領域に格納する。以下同様にData#19、22(D#19、22)と回復処理を行ない論理グループ10内のスペース領域に格納していく。

【0087】この様に、SD#2、3、4、5、6のSCSIドライブ12内のスペース領域に、SD#1のSCSIドライブ12の回復データを全て格納した後は、スペース領域が論理グループ10において一つしかないため、実施例1で述べたような書き込み時の回転待ちを短くすることは出来ないため、従来のアレイディスクであるRAIDのレベル5の処理となる。また、SD#1のSCSIドライブ12の回復データを全て格納した後は、SD#2、3、4、5、6のSCSIドライブ12の中で更にもう一台SCSIドライブ12に障害が発生した場合、同様にその障害が発生したSCSIドライブ12内のデータについて回復処理を行ない、残りのスペース領域に格納し、処理を行なえる。

【0088】この様にして、論理グループ10内のスペース領域を全て使いきってしまった場合は、障害SCSIドライブ12を正常のSCSIドライブ12に交換し、この交換した正常なSCSIドライブ12は全てスペース領域として論理グループを再構成する。

【0089】障害SCSIドライブ12を正常のSCSIドライブ12に交換した直後は、スペース領域が特定SCSIドライブ12に集中した形になっているため、このSCSIドライブ12が使用出来ずにまたされることが多くなりネックとなるため、実施例1で示した回転待ち時間を短縮する効果が、効率的に発揮出来ない。しかし、時間が立つにつれて、スペース領域が分散されてSCSIドライブ12障害前の状態に戻っていき、次第に解消されていく。もし、この時間が問題となる場合は、SCSIドライブ12に障害が発生したことを感知

した場合、正常なSCSIドライブ12に交換して、この交換した正常なSCSIドライブ12に障害が発生したSCSIドライブ12内のデータとパリティをユーザが復元することも可能とする。なお、この時スペース領域に関しては復元せずにスペース領域として空けておく。

【0090】本実施例ではこの回復処理と、スペース領域へ復元したデータを書き込む処理をMP120が独自に行なう。この様に独自に行なうことによりSCSIドライブ12に障害が発生した場合、障害が発生したSCSIドライブ12を正常なSCSIドライブ12に交換し回復したデータを書き込むのと比較し、本発明ではシステムを使用するユーザがSCSIドライブ12に障害が発生するとすぐに正常なSCSIドライブ12と交換する必要が無い場合、ユーザの負担が軽くなる。

【0091】〈実施例3〉本発明の第三の実施例を図7～図11により説明する。本実施例では図7、8に示すように論理グループ10単位にサブDKC11を設け、その内部に図9に示すように実施例1、2において示したキャッシュメモリ7内のアドレステーブル40とRPC27、PG36、サブキャッシュ32とそれらを制御するマイクロプロセッサMP329を持たせたものを示す。本実施例におけるデータの処理手順は実施例1および2で示したものと同様である。以下には実施例1、2で示した処理手順と異なる部分のみを図10および図11を用いて示す。本実施例では図9に示すように実施例1、2で示したキャッシュメモリ7内のアドレステーブル40をサブDKC11内のデータアドレステーブル(DAT)30に格納する。DAT30は格納されているテーブルの形式や機能は実施例1、2と同様であるが、異なるのはデータを格納するSCSIドライブアドレス42が論理グループ10に限られている点と、メモリがアドレステーブル40を格納するデータを格納するのは別の専用メモリである。ADC2内のGAT23はCPU1から指示されたCPU指定アドレスから、そのCPU指定アドレスが指示する場所がADU3内のどの論理グループ10かを判定するのみである。キャッシュメモリ7内には、その特別な領域に図10に示すような論理グループテーブル(LGT)60が格納されている。

【0092】LGT60は図10に示すようにCPU1から指定されるCPU指定ドライブ番号41とCCHHR46に対応して、論理グループアドレス61が決定できるテーブルとなっている。また、LGT60にはCPU指定アドレスに対応するデータがキャッシュメモリ7内に存在する場合、そのデータのキャッシュメモリ7内のアドレスをキャッシュアドレス47に登録でき、また、キャッシュメモリ7内にデータが存在する場合オン(1)とし、キャッシュメモリ7内に存在しない場合オフ(0)とするキャッシュフラグ48が用意されている。

る。ユーザは初期設定する際に自分の使用可能な容量に対する領域を確保するが、その際にADC2のMP120が、LGT60により論理グループ10を割当てる。この時、MP120はLGT60にユーザが確保するために指定したCPU指定アドレスに対応する領域を登録する。

【0093】そこで、実際の読みだし、書き込み処理においては、GAT23はLGT60によりCPU1から指定してきたCPU指定アドレスに対応した論理グループ10を認識することが可能となる。読み出し時はGAT23がLGCにより論理グループ10を確定し、その確定した結果をMP120に報告し、MP120はこの当該論理グループ10に対し読み出し要求を発行するようにDriveIF28に指示する。MP120から指示を受けたDriveIF28は当該論理グループ10のサブDKC11に対し読み出し要求とCPU1が指定するCPU指定アドレスを発行する。サブDKC11ではマイクロプロセッサであるMP329がこの読み出し要求のコマンドとCPU指定アドレスを受け付け、実施例1と同様に、DriveIF28から送られてきたCPU指定アドレスをDAT30を参照し、当該データが格納されている論理グループ10内の論理アドレス45に変換し、この論理アドレス45から当該SCSIドライブアドレス42（SCSIドライブ番号43とその中のSCSI内Addr44）を確定する。

【0094】このアドレスの確定後MP329は当該SCSIドライブ12に対し、読み出し要求を発行する。MP329から読み出し要求を発行されたSCSIドライブ12ではSCSI内Addr44ヘシーク、回転待ちを行ない、当該データの読み出しが可能になり次第、当該データをドライブアダプタ回路（DriveAdp）34に転送し、DriveAdp34はサブキャッシュメモリ32に格納する。サブキャッシュメモリ32に当該データの格納が完了した後、DriveAdp34はMP329に格納報告を行ない、MP329はDAT30の当該データの当該論理アドレス45内の当該キャッシュフラグ48をオン（1）とする。後に当該キャッシュフラグ48がオン（1）のデータに対し読み出しまたは書き込み要求が発行された場合は、サブキャッシュ32内で処理を行なう。実施例1と同様にMP329によるDAT30の更新が終了すると、MP329はADC2内のDriveIF28に対しデータ転送可能という応答を行ない、DriveIF28はこの応答を受け取ると、MP120に対し報告する。

【0095】MP120はこの報告を受け取ると、キャッシュメモリ7への格納が可能なら、DriveIF28に対しサブDKC11からデータを転送するように指示する。DriveIF28ではMP120からの指示を受けるとサブDKC11のMP329に対

し読み出し要求を発行する。この読み出し要求を受けたMP329はサブキャッシュアダプタ回路（SCA）31に対しサブキャッシュ32から当該データを読みだすように指示し、SCA31は実際にデータを読み出してDriveIF28にデータを転送する。DriveIF28がデータを受け取った後は、実施例1、2で示した処理を行なう。

【0096】一方書き込み時は読み出し時と同様に当該論理グループ10を確定し、MP120はDriveIF28に対し当該論理グループ10のMP329に対し書き込み要求を発行するように指示する。当該論理グループ10内のMP329は、書き込み要求を受け付け、書き込みデータをサブキャッシュ32に格納した後は、図5のフローチャートに従って、実施例1、2と同様に処理を行う。本実施例では実施例1、2の効果を実現することが可能である。

【0097】以上、磁気ディスク装置を用いたシステムを実施例として説明したが、本発明は光ディスク装置を用いたシステムにおいても同様な効果を発揮することが可能である。

【0098】

【発明の効果】本発明によれば、データの書き込み時におけるパリティの更新処理をCPUからの読みだしまたは書き込み要求が少ない時まで遅らせることが可能となる。これにより、CPUにとっては読みだしまたは書き込み処理要求が多い時は書き込み処理を高速に行え、これにより単位時間当りのI/O処理件数を増加させることが可能となる。さらに、通常は使用しないスベアのSCSIドライブを、回転待ち時間の短縮という、性能向上のために使用でき、SCSIドライブ資源の有効活用が図れる。

【図面の簡単な説明】

【図1】第1の実施例の全体構成図。

【図2】第1の実施例のクラスタ内構成図。

【図3】アドレス変換テーブルの説明図。

【図4】書き込み処理時のデータ移動説明図。

【図5】書き込み処理フローチャート（1）。

【図6】データ回復処理説明図、パリティグループを構成する各データのディスク上での位置説明図。

【図7】第3の実施例の全体構成図。

【図8】第3の実施例のクラスタ内構成図。

【図9】第3の実施例のサブDKC内構成図。

【図10】論理グループテーブル説明図。

【図11】RAID Level 5における更新処理説明図。

【図12】RAID 5における書き込み処理タイミングチャート。

【図13】書き込み処理フローチャート（2）。

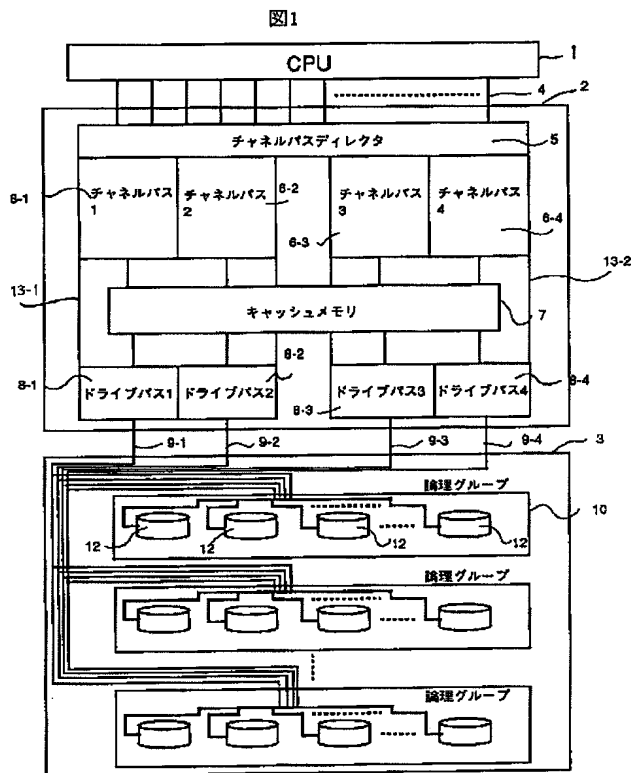
【図14】パリティ作成処理フローチャート。

【符号の説明】

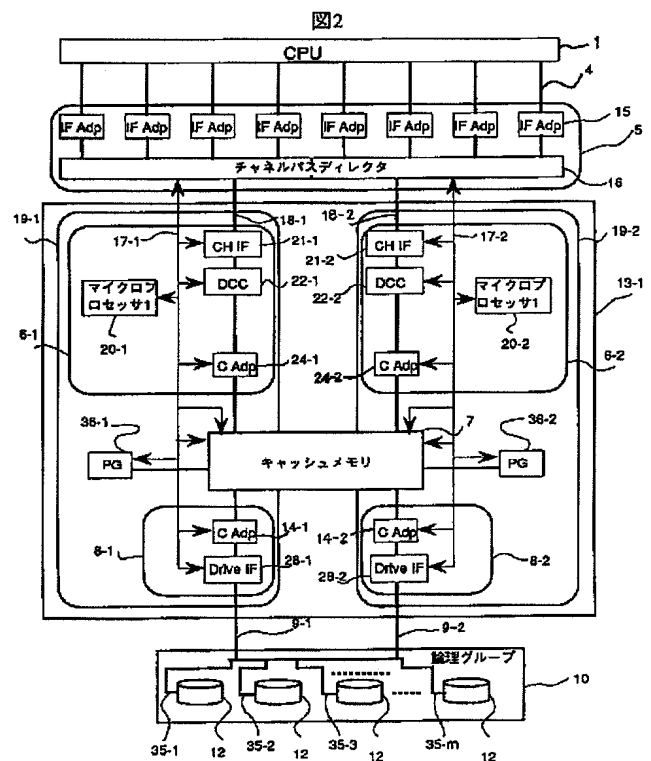
1…CPU、2…アレイディスクコントローラ(ADC)、3…アレイディスクユニット(ADU)、4…外部インターフェースバス、5…チャンネルバスディレクタ、6…チャンネルバス、7…キャッシュメモリ、8…ドライブバス、9…アレイディスクユニットバス、10…論理グループ、12…SCSIドライブ、13…クラスタ、14…ドライブ側キャッシュアダプタ(C Adp)、15…インターフェースアダプタ、16…チャンネルバススイッチ、17…制御信号線、18…データ線、19…パス、20…マイクロプロセッサ1(MP1)、21…チャンネルインターフェース(CH IF)回路、22…データ制御回路(DCC)、23…グループアドレス変換回路(GAT)、24…チャンネル側キャッシュアダプタ(C Adp)、28…ドライブインターフェース*

*ース回路(Drive IF)、29…マイクロプロセッサ3(MP3)、30…データアドレステーブル40、31…サブキャッシュアダプタ、34…ドライブアダプタ(Drive Adp)、35…ドライブバス、36…パリティ生成回路、40…アドレス変換用テーブル(アドレステーブル)、41…CPU指定ドライブ番号、42…SCSIドライブアドレス、43…SCSIドライブ番号、44…SCSI内Addr、45…論理アドレス、46…CCHHR、47…キャッシュアドレス、48…キャッシュフラグ、49…無効フラグ、50…パリティドライブ番号、51…スペースドライブ番号、52…ドライブフラグ、53…SDフラグ、60…論理グループテーブル、61…論理グループアドレス

【図1】



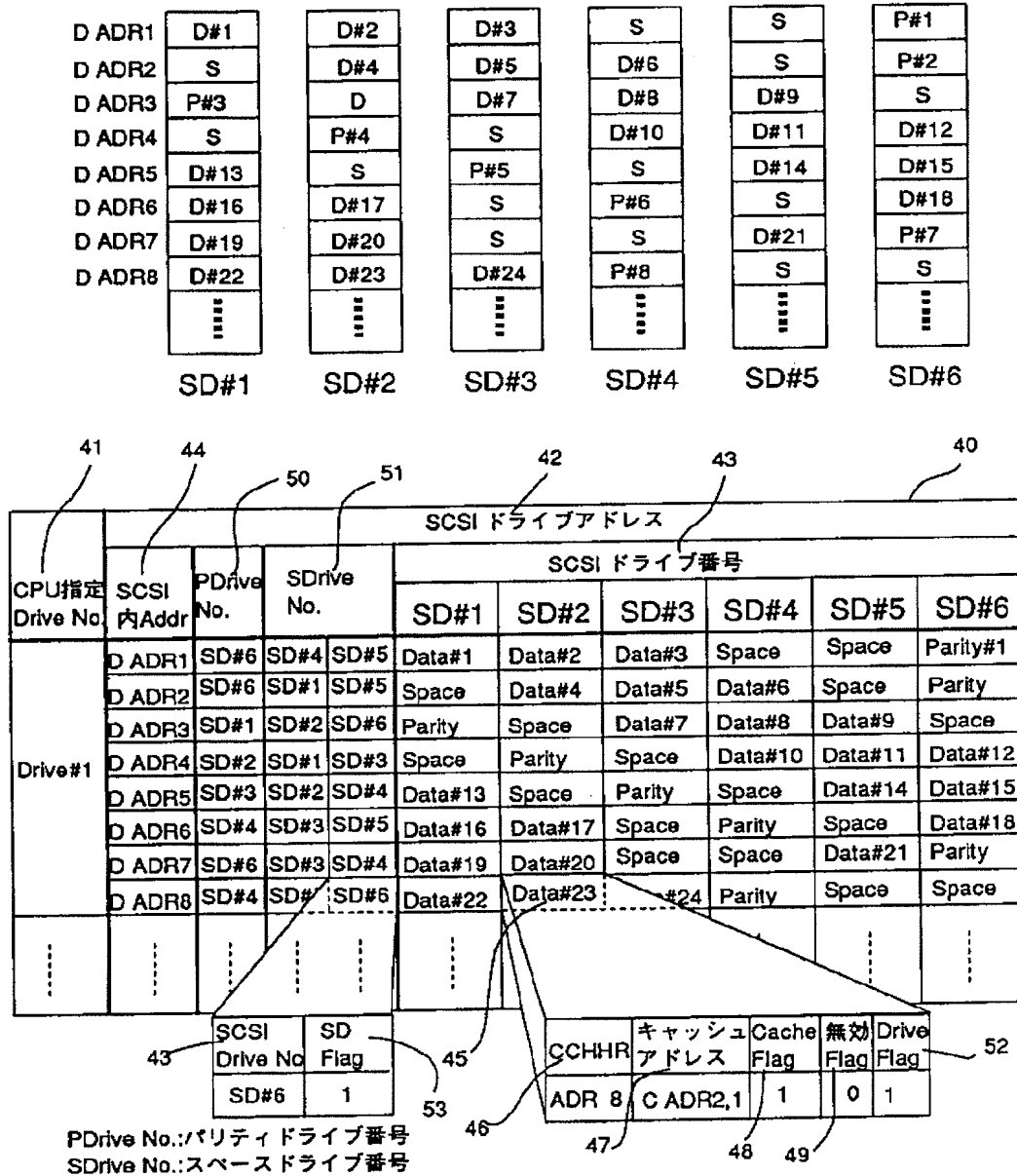
【図2】



【図3】

図3

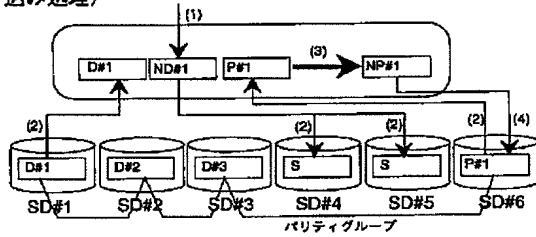
D: データ P: パリティ S: スペース領域



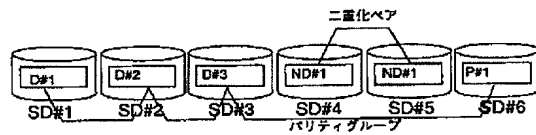
【図4】

図4

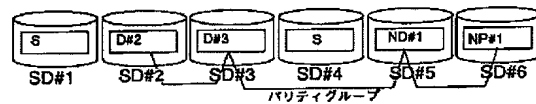
(書き込み処理)



(a) (4)のSD#6のSCSIドライブへのパリティ(NParity#1)の書き込み前

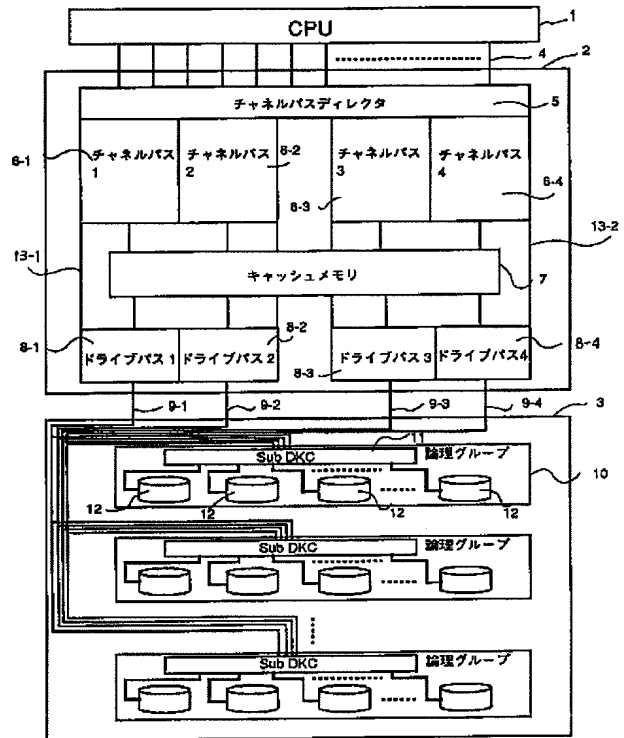


(b) (4)のSD#6のSCSIドライブへのパリティ(NParity#1)の書き込み後



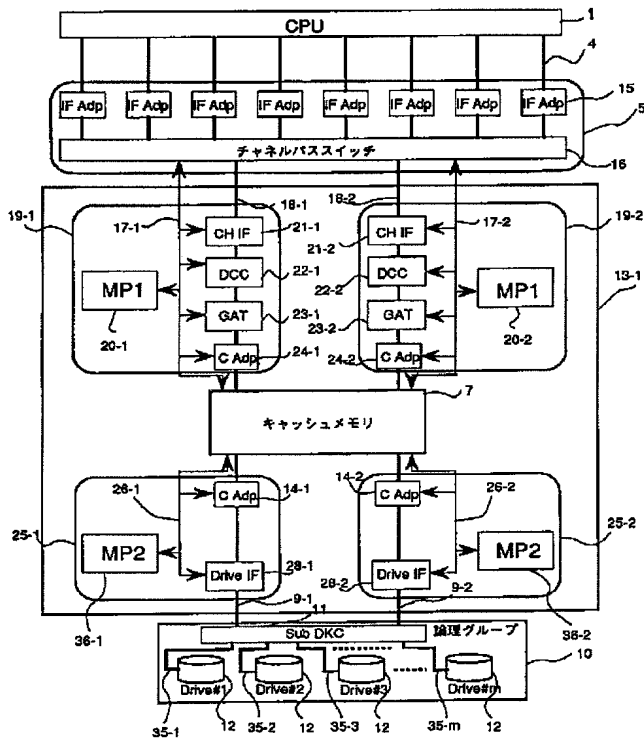
【図7】

図7



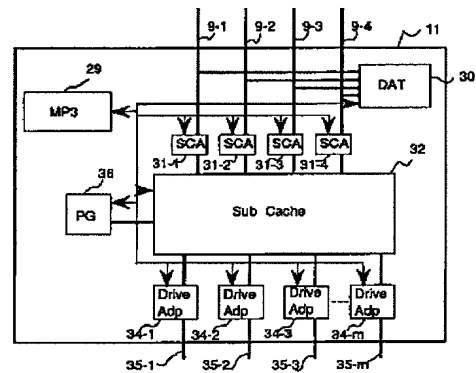
【図8】

図8

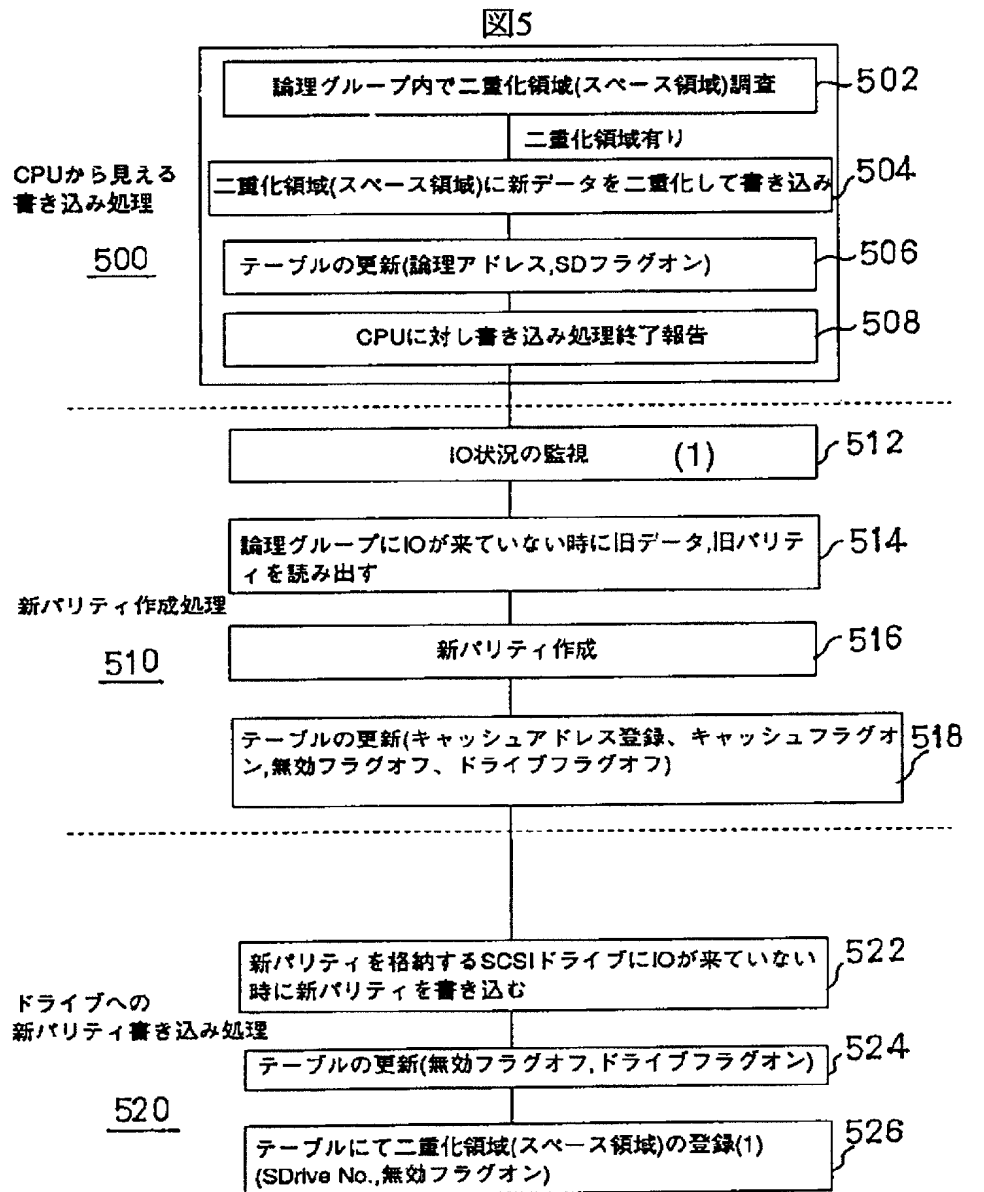


【図9】

図9



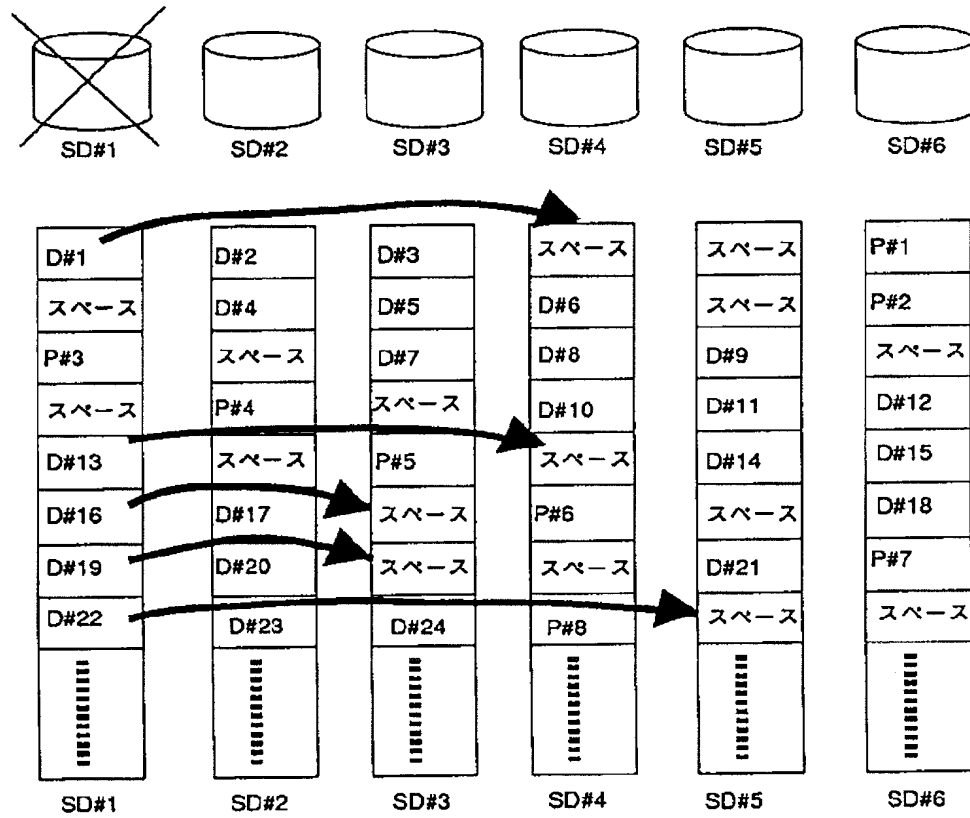
【図5】



(1)二重化領域は書き込み先に指定したアドレスに書き込まれているデータと二重化されているデータのペアの中でSCSIドライブ番号の小さい方とする

【図6】

図6

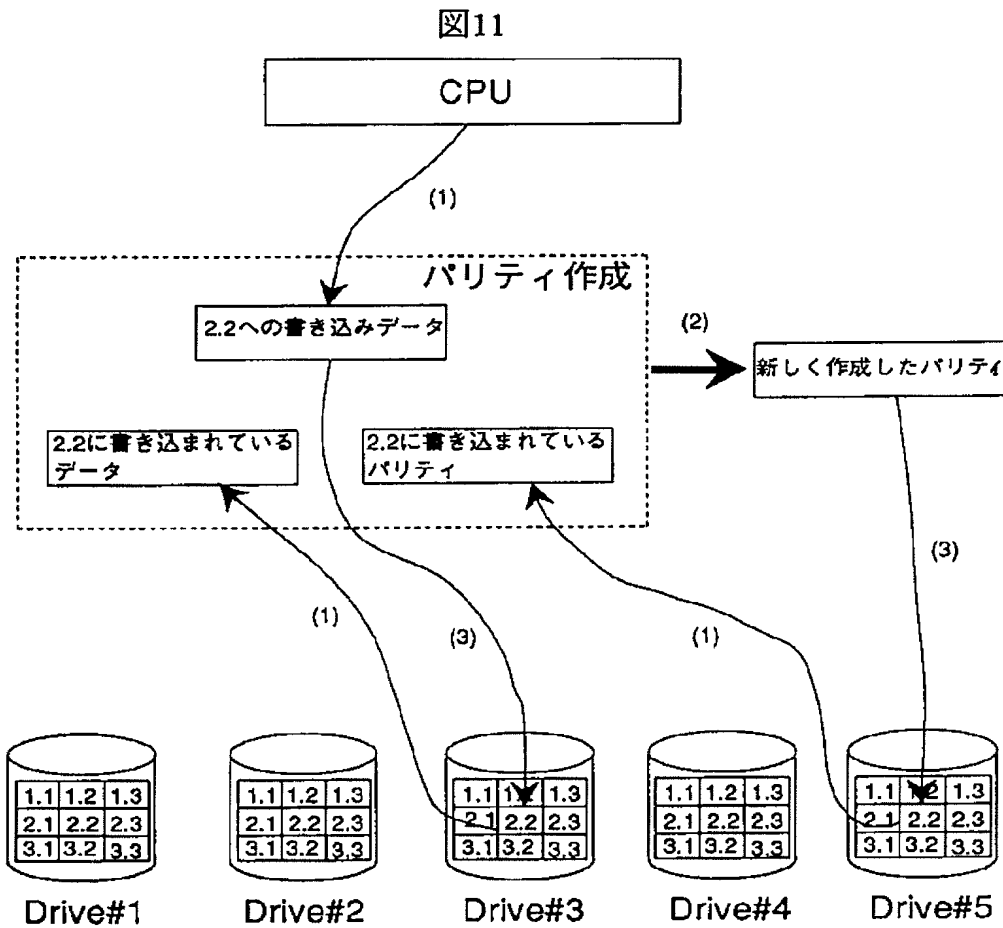


【図10】

図10

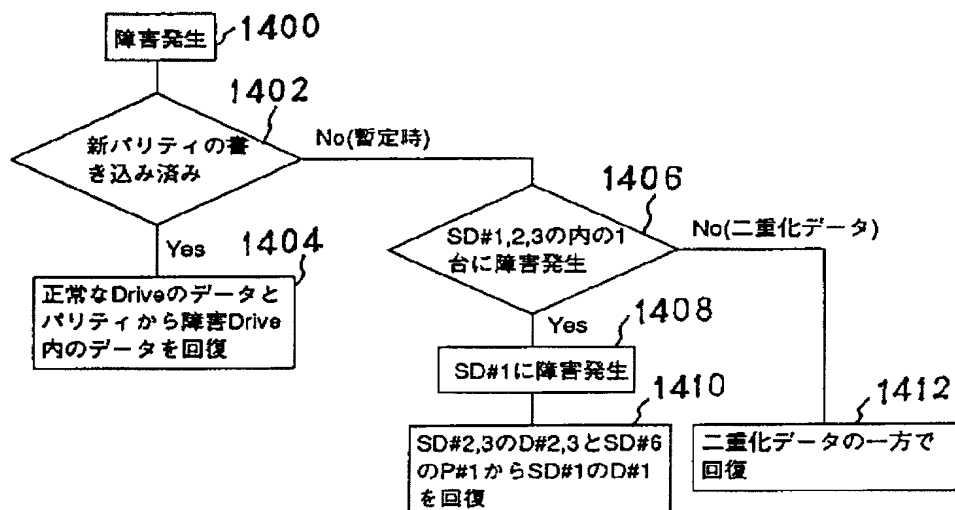
GPU指定アドレス		46	61	47	60	48
CPU指定 Drive No.	CCHHR	論理グループ アドレス	キャッシュ アドレス	Cache Flag		
Drive#1	ADR 1	LADR 1	---	---		
	ADR 2	LADR 3	---	---		
	ADR 3	LADR 8	---	---		
	⋮	⋮	⋮	⋮		
Drive#2	ADR 1	LADR 2	---	0		
	ADR 2	LADR 1	CADR1,5	1		
	ADR 3	LADR 5	CADR1,8	1		
	ADR 4	LADR 4	CADR1,8	1		
⋮	⋮	⋮	⋮	⋮		

【図11】



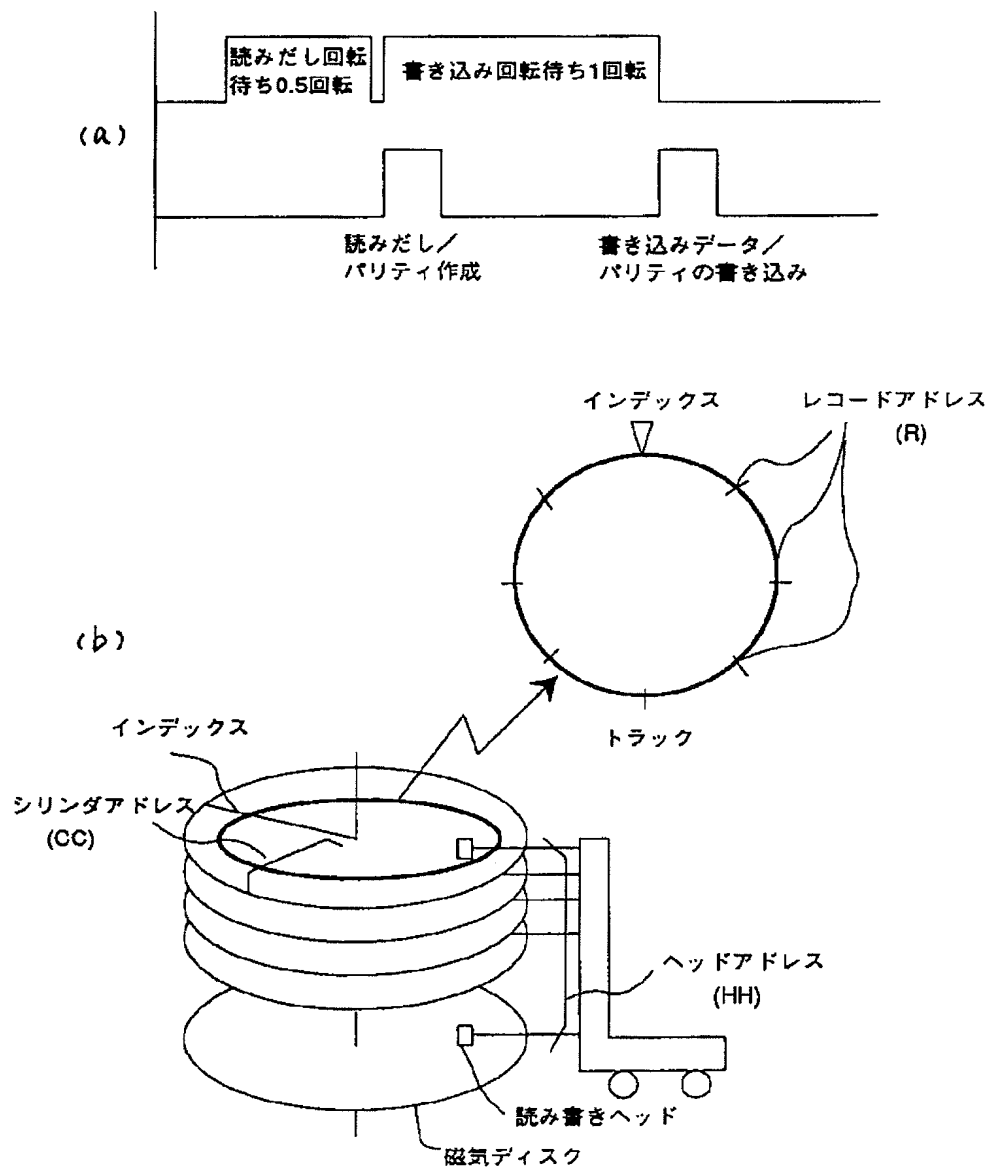
【図14】

図14



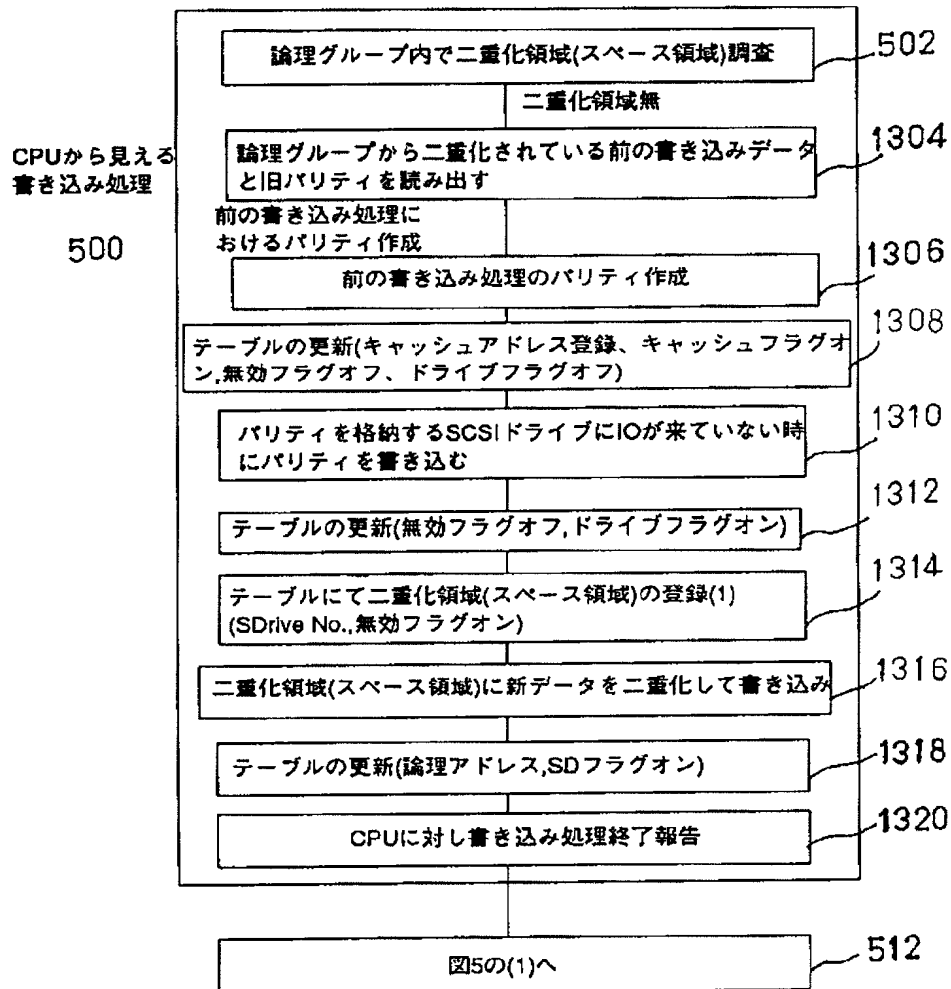
【図12】

図12



【図13】

図13



(1)二重化領域は書き込み先に指定したアドレスに書き込まれているデータと二重化されているデータのペアの中でSCSIドライブ番号の小さい方とする

【公報種別】特許法第17条の2の規定による補正の掲載
 【部門区分】第6部門第3区分
 【発行日】平成13年2月9日(2001.2.9)

【公開番号】特開平6-332632
 【公開日】平成6年12月2日(1994.12.2)
 【年通号数】公開特許公報6-3327
 【出願番号】特願平5-125766
 【国際特許分類第7版】

G06F 3/06 305
 301
 11/10 320

【F I】

G06F 3/06 305 C
 301 Z
 11/10 320 Z

【手続補正書】

【提出日】平成12年4月14日(2000.4.14)

【手続補正1】

【補正対象書類名】明細書

【補正対象項目名】特許請求の範囲

【補正方法】変更

【補正内容】

【特許請求の範囲】

【請求項1】データを格納する複数のドライブ群と、該ドライブ群を管理する制御装置とを備え、書き込みデータを前記ドライブ群に格納し、該各ドライブに格納されているデータにより誤り訂正符号を生成し、この生成した誤り訂正符号を、該誤り訂正符号を生成するのに関与したデータが格納されているドライブとは別のドライブに格納する記憶装置の制御方法において、データ書き込みに伴う誤り訂正符号の更新を、前記データ書き込み処理よりも遅延させて処理することを特徴とする記憶装置の制御方法。

【請求項2】書き込むべき複数のデータから少なくとも1つの誤り訂正符号を含む1つの論理グループを生成し、該論理グループを構成する複数のデータ及び誤り訂正符号を複数のドライブ群内に格納する記憶装置におけるデータの格納方法であって、前記論理グループを構成する複数のデータの1つを新たなデータに書き換える更新要求に応答して、該更新データと新たな誤り訂正符号とを含む1つの論理グループを新たに生成し、前記新たな論理グループの更新データと誤り訂正符号とを、前記書き換え前のデータ及び誤り訂正符号が格納されていたドライブとは異なるドライブにそれぞれ分散して格納する記憶装置におけるデータの格納方法において、データ書き込みに伴う誤り訂正符号の更新を、前記データ書き込み処理よりも遅延させて処理することを特徴と

する記憶装置の制御方法。

【請求項3】データを格納する複数のドライブ群と、該ドライブ群を管理する制御装置とを備え、書き込みデータを前記ドライブ群に格納し、該各ドライブに格納されているデータにより誤り訂正符号を生成し、この生成した誤り訂正符号を、該誤り訂正符号を生成するのに関与したデータが格納されているドライブとは別のドライブに格納する記憶装置の制御方法において、

データ書き込みに伴う誤り訂正符号の更新を、データの書き込みを要求した上位置装置の負荷が所定値以下のときに処理することを特徴とする記憶装置の制御方法。

【請求項4】書き込むべき複数のデータから少なくとも1つの誤り訂正符号を含む1つの論理グループを生成し、該論理グループを構成する複数のデータ及び誤り訂正符号を複数のドライブ群内に格納する記憶装置におけるデータの格納方法であって、前記論理グループを構成する複数のデータの1つを新たなデータに書き換える更新要求に応答して、該更新データと新たな誤り訂正符号とを含む1つの論理グループを新たに生成し、前記新たな論理グループの更新データと誤り訂正符号とを、前記書き換え前のデータ及び誤り訂正符号が格納されていたドライブとは異なるドライブにそれぞれ分散して格納する記憶装置におけるデータの格納方法において、データ書き込みに伴う誤り訂正符号の更新を、データの書き込みを要求した上位置装置の負荷が所定値以下のときに処理することを特徴とする記憶装置の制御方法。

【請求項5】データを格納する複数のドライブ群と、該ドライブ群を管理する制御装置とを備え、書き込みデータを前記ドライブ群に格納し、該各ドライブに格納されているデータにより誤り訂正符号を生成し、この生成した誤り訂正符号を、該誤り訂正符号を生成するのに関与したデータが格納されているドライブとは別のドライブ

に格納するにおいて、

データ書き込みに伴う誤り訂正符号の更新を、前記制御装置の負荷が所定値以下のときに処理することを特徴とする記憶装置の制御方法。

【請求項6】書き込むべき複数のデータから少なくとも1つの誤り訂正符号を含む1つの論理グループを生成し、該論理グループを構成する複数のデータ及び誤り訂正符号を複数のドライブ群内に格納する記憶装置におけるデータの格納方法であって、前記論理グループを構成する複数のデータの1つを新たなデータに書き換える更新要求にตอบสนองして、該更新データと新たな誤り訂正符号とを含む1つの論理グループを新たに生成し、前記新たな論理グループの更新データと誤り訂正符号とを、前記書き換え前のデータ及び誤り訂正符号が格納されていたドライブとは異なるドライブにそれぞれ分散して格納する記憶装置におけるデータの格納方法において、データ書き込みに伴う誤り訂正符号の更新を、前記制御装置の負荷が所定値以下のときに処理することを特徴とする記憶装置の制御方法。

【請求項7】データを格納する複数のドライブ群と、該ドライブ群を管理する制御装置とを備え、書き込みデータを前記ドライブ群に格納し、該各ドライブに格納されているデータにより誤り訂正符号を生成し、この生成した誤り訂正符号を、該誤り訂正符号を生成するのに関与したデータが格納されているドライブとは別のドライブに格納するディスクレイシステムにおいて、データ書き込みに伴う誤り訂正符号の更新を、前記データの書き込みとは非同期でかつ所定のタイミング毎に処理することを特徴とするディスクレイ装置の制御方法。

【請求項8】あるドライブに障害が発生し、障害が発生したドライブ内のデータまたはパリティを回復処理により復元し、この復元したデータまたはパリティを予備の書き込み領域に書き込んだ後、障害が発生したドライブを、正常なドライブに交換し、交換後は、この交換した正常なドライブは全てスペース領域により構成されているとして論理グループを再構成して処理を再開することを特徴とする請求項1記載の記憶装置の制御方法。

【請求項9】あるドライブに障害が発生したことを感知したら、障害が発生したドライブを正常なドライブに交換し、障害が発生したドライブ内のデータまたはパリティを回復処理により復元し、この復元したデータまたはパリティと、障害が発生したドライブ内にあったスペース領域を、交換した正常なドライブに書き込んで再構成して処理を再開することを特徴とする請求項1記載の記憶装置の制御方法。

【請求項10】最新に書き込まれたデータについては二重化して高信頼とし、書き込み要求があまり発行されないデータについてはパリティにより信頼性を確保するように、信頼性について二段階のレベルを設定することを

特徴とする請求項1記載の記憶装置の制御方法。

【請求項11】書き込むべき複数のデータから少なくとも1つの誤り訂正符号を含む1つの論理グループを生成し、該論理グループを構成する複数のデータ及び誤り訂正符号を複数のドライブ群内に格納する記憶装置であって、前記論理グループを構成する複数のデータの1つを新たなデータに書き換える更新要求にตอบสนองして、該更新データと新たな誤り訂正符号とを含む1つの論理グループを新たに生成し、前記新たな論理グループの更新データと誤り訂正符号とを、前記書き換え前のデータ及び誤り訂正符号が格納されていたドライブとは異なるドライブにそれぞれ分散して格納する記憶装置において、データ書き込みに伴う誤り訂正符号の更新を、前記データ書き込み処理よりも遅延させて処理することを特徴とする記憶装置。

【請求項12】前記記憶装置を構成するドライブの集合の中に、書き込むデータを一旦二重化して書き込める領域をもつことを特徴とする請求項11記載の記憶装置。

【請求項13】スペース領域を、パリティを生成するデータと、それらのデータにより生成したパリティが格納されているドライブに分散してもつことを特徴とする請求項11記載の記憶装置。

【請求項14】スペース領域を、パリティおよびパリティの作成に関与したデータの格納されているドライブ以外の、異なる2台のドライブに確保することを特徴とする請求項11記載の記憶装置。

【請求項15】論理グループにおいて、パリティを作成するデータとパリティの集合をパリティグループとし、書き込み前と書き込み後では、パリティグループを構成するデータおよびパリティの格納されているドライブが異なることを特徴とする請求項11記載の記憶装置。

【請求項16】論理グループ内のあるドライブに障害が発生した場合、正常な残りのドライブ内のデータとパリティから、障害が発生したドライブ内のデータまたはパリティを回復処理により復元するが、この復元したデータまたはパリティをスペース領域に書き込む制御を行うプロセッサを持つことを特徴とする請求項11記載のディスクレイシステム。

【請求項17】書き込みにより作成した新パリティを、当該ドライブへ書き込む前に、あるドライブに障害が発生した場合、旧パリティの作成に関与したデータについては、この旧パリティの作成に関与した、正常な残りのドライブ内のデータとパリティから回復処理により復元し、二重化されている新データが格納されているドライブに障害が発生した場合は、二重化データの一方から、障害が発生したドライブ内のデータまたはパリティを回復処理により復元する制御を行うプロセッサを持つことを特徴とする請求項11記載のディスクレイシステム。

【請求項18】書き込みにより作成した新パリティを、

当該ドライブへ書き込む前に、あるドライブに障害が発生した場合、正常な残りのドライブ内のデータとパリティと二重化されている新データから、障害が発生したドライブ内のデータまたはパリティを回復処理により復元するが、この復元したデータまたはパリティをスペース領域に書き込む制御を行うプロセッサを持つことを特徴とする請求項1記載のディスクアレイシステム。

【請求項19】外部装置から入力されたデータを記憶装置に書き込む方法であって、

入力データを第1の冗長レベルの形式で上記記憶装置に書き込むステップと、

上記記憶装置に第1の冗長レベルの形式で書き込まれた上記データを第2の冗長レベル形式に書き換えるステップとを有することを特徴とするデータ書き込み方法。

【請求項20】前記外部装置と前記記憶装置との間のデータの入力または出力の頻度を検出するステップを有し、

上記検出された頻度に応じて、前記第1の冗長レベルの形式で書き込まれたデータを第2の冗長レベルの形式に書き換えることを特徴とする請求項19記載のデータ書き込み方法。

【請求項21】前記第1の冗長レベルの形式が2重化を

利用するものであり、前期第2の冗長レベルの形式がパリティを利用するものであることを特徴とする請求項19または請求項20に記載のデータ書き込み方法。

【請求項22】外部装置と接続可能であって、上記外部装置からの入力データを書き込み、内部に記録されているデータを上記外部装置に読み出す記憶装置であって、上記外部装置からの入力データを第1の冗長レベルの形式で書き込む手段と、

上記第1の冗長レベルの形式で書き込まれたデータを第2の冗長レベル形式に書き換える手段とを有することを特徴とする記憶装置。

【請求項23】前記外部装置と前記記憶装置との間のデータの入力または出力の頻度を検出する手段を有し、前記書き換え手段が、上記検出手段によって検出された頻度に応じて、前記第1の冗長レベルの形式で書き込まれたデータを第2の冗長レベルの形式に書き換えることを特徴とする請求項22記載の記憶装置。

【請求項24】前記第1の冗長レベルの形式が2重化を利用するものであり、前期第2の冗長レベルの形式がパリティを利用するものであることを特徴とする請求項22または請求項23に記載の記憶装置。